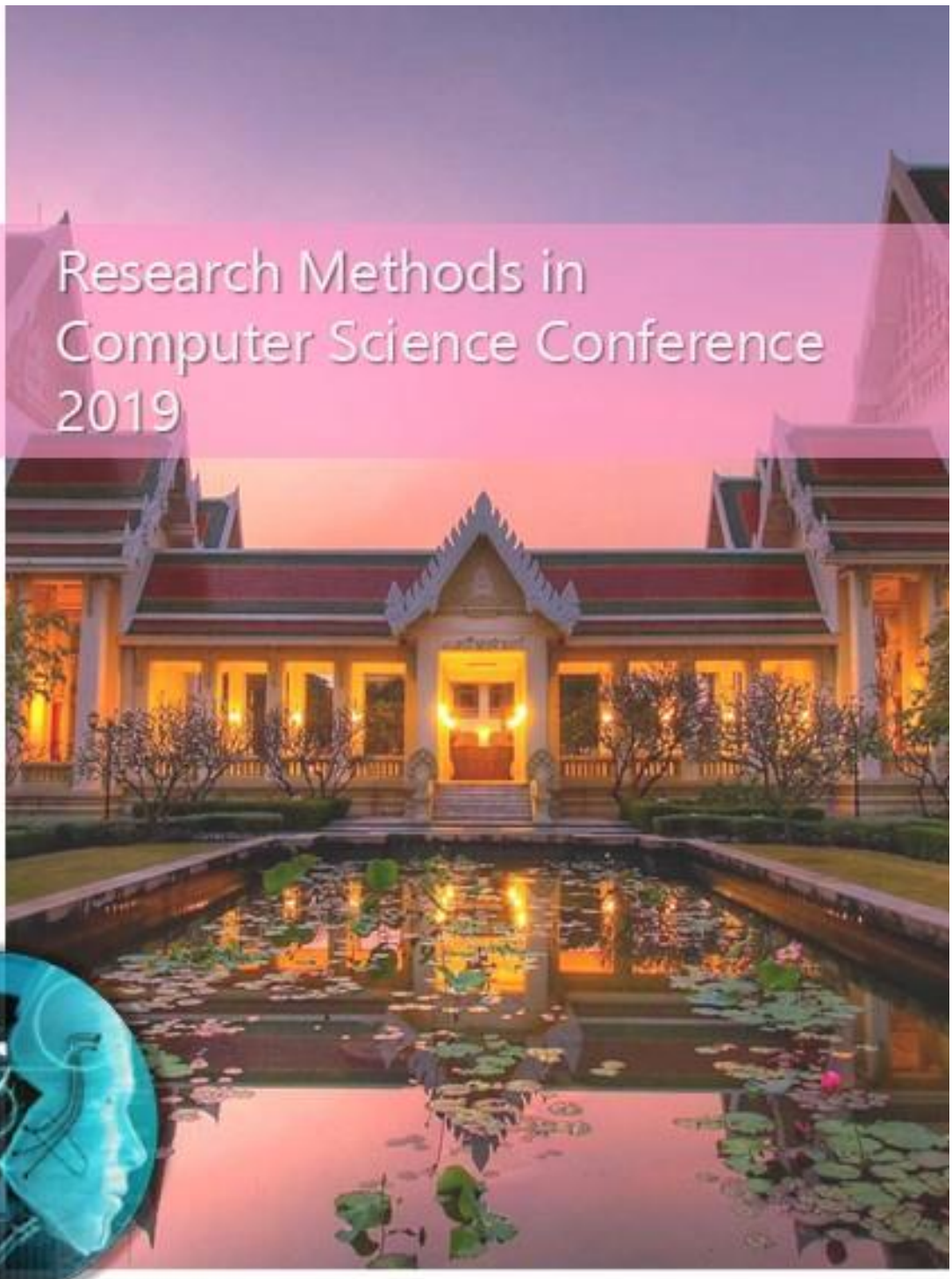
Research Methods in Computer Science Conference 2019



1 December 2019

Meeting room @19th floor
Engineering 4 Building (Charoenvidsavakham)
Chulalongkorn University

Computer Engineering
Chulalongkorn University

การเรียนรายวิชา Research Methodology มีวัตถุประสงค์มุ่งเน้นให้นิสิตมีความรู้ความเข้าใจในกระบวนการทำงานวิจัย ตลอดจนการวิเคราะห์และสรุปผลการวิจัยเพื่อเขียนบทความวิชาการ และมีประสบการณ์ในการนำเสนองานวิชาการและเผยแพร่ผลงานในวารสารวิชาการได้

การจัดงานสัมมนาทางวิชาการ Research Methods in Computer Science Conference 2019 ครั้งนี้ จึงทำให้นิสิตเข้าใจกระบวนการทำงานวิจัยอย่างสมบูรณ์ยิ่งขึ้น

ขอขอบคุณ ศ.ดร. ประภาส จงสถิตย์วัฒนา ที่สนับสนุนการดำเนินงานตลอดจนขอบคุณนิสิตปริญญาเอกในห้องปฏิบัติการ Intelligent System Laboratory ที่ช่วยกันตรวจทานและให้คำแนะนำในการเขียนบทความวิชาการ เพื่อให้บทความวิชาการที่ได้มีความถูกต้องสมบูรณ์มากยิ่งขึ้น และขอบคุณนิสิตทุกท่านที่ลงทะเบียนเรียนวิชา 2110607 Research Methods in Computer Science ในภาคการศึกษาต้น ปีการศึกษา 2562 ที่ช่วยกันสนับสนุนให้งานสัมมนาในครั้งนี้เป็นไปได้อย่างสมบูรณ์

Program

At Meeting room 19th floor, Engineering 4 Building (Charoenvidsavakham),
Department of Computer Engineering, Faculty of Engineering,
Chulalongkorn University

Time	Program
12.45	Registration
13.00	Introductory Talk by Prof.Dr. Prabhas Chongstitvatana
13.10	Invited Talk: Break RSA encryption by Shor's Algorithm Mr. Wiphoo Methachawalit Refinitiv Software (Thailand)
13.40	Food image retrieval and recognition using Convolutional Neural Network
13.55	Extract keywords embedded in comments from YouTube to predict trends
14.10	Cashflow prediction of Automated Teller Machine in Bangkok
14.25	Violence detection from surveillance camera
14.40	YouTube video classification of violent content based on comments
14.55	Analysis of touristic places in Thailand from Facebook
15.10	Break
15.30	Sentiment Analysis of Twitter data to predict credibility of online shops
15.45	Prediction of the success and the number of goal in football game
16.00	Aspect-Level sentiment analysis on movies review using Deep Learning
16.15	Eye blink detection using machine learning for prevention of computer vision syndrome
16.30	Detecting a sign of Major Depressive Disorder from social network activities
16.45	Closing

Table of Content

Break RSA encrpytion by Shor's algorithm

Wiphoo Methachawalit

Shor's algorithm is the one of the most famous algorithm on Quantum computer. This is an algorithm proposed the way to break RSA encryption in 1997. In recent year, quantum computers are now more available for researchers. The superconducting quantum computers are developing and investing by big name companies like IBM, Google and Rigetti.

IBM Q experience is quantum computing that open for researchers to develop their application or experimenting with quantum computer. The RSA encryption will be cracked by Shor's algorithm using currently superconducting quantum computer.

Biography:

Wiphoo Methachawalit received B.S. degree in Computer Engineering from King Mongkut's University of Technology Thonburi, Thailand, in 2010. He experienced as Head Research Development Department with a demonstrated history of working in the media production industry. Skilled in Python, C++, Databases, OpenGL, and Software Documentation. Currently working with Refinitiv Software (Thailand) as Lead Software Engineer.

Food image retrieval and recognition using Convolutional Neural Network

Apichad Chodkawanich, Jirapong Poolsuk, Pantira Jantanawaranon
Department of Computer Engineering
Faculty of Engineering
Chulalongkorn University
Bangkok, Thailand
apichad.chod@gmail.com, Jirapong.birth@gmail.com, pantira-jan@hotmail.com

***Abstract*—We propose a food image recognition system with Convolutional Neural Networks (CNN). CNN has been applied to image recognition successfully in the literature. The network consists of four layers has been built. Ten dataset are prepared from Wongnai website. We have achieved the best accuracy of 71.02% on the dataset.**

Keywords—Convolutional Neural Network, food photography, Image recognition

I. INTRODUCTION

During recent decades, the role of the media - especially the social media sites in our lives have grown significantly, and their influence on culture and society is now extensive. Common activities on THE SOCIAL MEDIA SITES are viewing and posting pictures of tempting food that relies heavily on reviews and recommendations – such as Facebook, Instagram, Twitter, Pantip, Wongnai and etc.

Apart from food, there are numbers of new cafes that are trending in THE SOCIAL MEDIA SITES. There are too many for people to follow, especially in the city center like Siam Paragon or Siam Square One. Due to the shift toward social sharing, we often see friends' stories or news feed posts of eye-catching desserts that can promote impulse cravings. Unfortunately, most of the time, these restaurant locations are not shared. Thus, we propose an application, which users can upload the pictures of the dish and find out from which restaurant it is from.

II. RELATED WORK

Image analysis has the most popular feature called Convolutional Neural Networks. Both [1] Bag of color features (BoCF) and Bag of texture features (BoTF) are number of methods have been proposed by extract the histogram based on the learned color and texture, also used combined BoCF and BoTF and finally classify the food images with a linear support vector machine (SVM) classifier. The other applied deep learning, which was a popular method recently in the image recognition field. Weishan Zhang et al. [2] proposed four different datasets. The model achieved an accuracy of 80.8% for the fruit image dataset. In 2018, Ukrit Tiankaew et al.[3] built a new CNN model architecture using transfer learning in the modified VGG19 with 13 different kinds of Thai local food (7,632 total images after clean the dataset). The model achieved 82% test accuracy.

III. METHOD

Our model consists of 4-layer sub-blocks with 224×224 image input and the first layer having (32×3×3) feature map. The remaining layers compose of (64×3×3), (128×3×3), and (256×3×3) sizes respectively with each sub-layer block having pooling. The final block which is fully connected containing dropout. In summary, the network has 9,828,426 total parameters.

A. Building a dataset using web scraping

In the process of getting and preparing data, we use Python and BeautifulSoup to write a script that search for images of each class on Wongnai. The brief fifth steps have outlined [4]. The first is about connecting to a webpage. The second step is method of parsing html using BeautifulSoup. The third operates looping through the soup object to find elements of dataset. The next step shows performance of some simple data cleaning. At the Last step creates the data to a CSV file.

B. Convolutional neural networks

In the last few years, CNN has been widely used in food recognition applications, and it achieved better performance than the conventional machine learning methods.

CNN's are typically a configuration of three types of layers. Convolutional Layer, Subsampling Layer and Fully Connected Layer.

- The convolution layer takes the convolution of the input image with the convolution matrix and generates the output image. Usually, the convolution matrix is called filter and the output image is called filter response or filter map.
- This layer is mainly to reduce the size of the feature map for a faster processing time. A widely used pooling algorithm is max pooling. It extracts sub-sections of the feature map finds their maximum value, and drops all the other values
- This layer perform classification on the extracted features after the down sampling process by the pooling layers

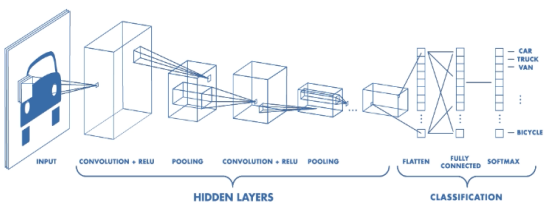


Fig. 1. Example convolutional neural network architecture

C. *Food retrieval*

First step for retrieving input images. We list 10 bakery shops' links around Siam Paragon from website www.wongnai.com into CSV file. Then we run our Python program for collecting list of photo's URL and photo's name from elements of Wongnai's website inserts into bakery shop's name CSV file. In the python program has processed getting 200 photos' data elements of each bakery shop by using requests library of python programing language.

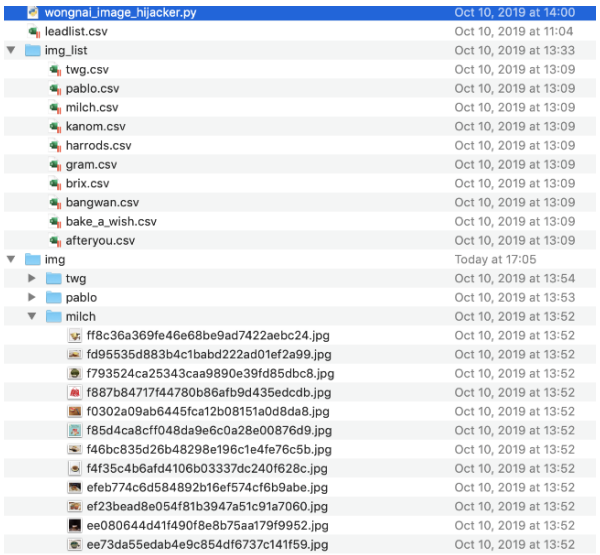


Fig. 2. Example images from Wongnai's website

IV. EXPERIMENTAL ANALYSIS

A. *Dataset*

To demonstrate the validation accuracy of our proposed model, we used the dataset which we get from Wongnai website. The dataset consists of 2,000 dessert images of 10 restaurant including: After you, Bake a wish, Bangwan, Brix, Gram, Harrods, Kanom, Milch, Pablo and Twg. 70 percent of the images were allocated to training set and the remaining 30 to validation set.

Fig. 3. Example dataset

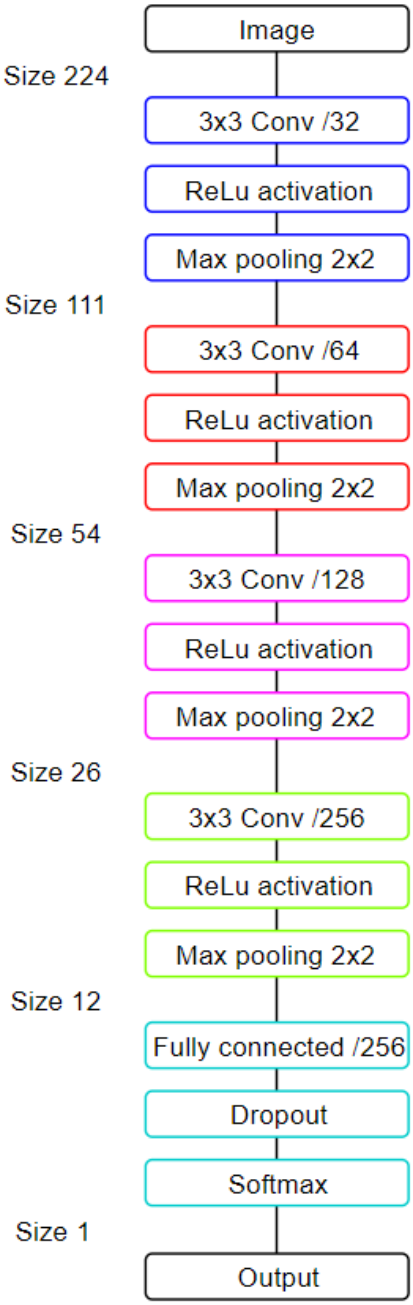


Fig. 4. Proposed CNN architecture

B. *Result*

A lot of technique like dropout were employed to fit our small dataset into our large model. We solely used 30 epochs in the entire experiment. In addition, we adopted smaller batch size to improve the classification performance of our model although with higher computation demand and time. In figure 3, we obtained 93.02% training accuracy and 71.02% test accuracy with training size of 224×224 on both dataset.

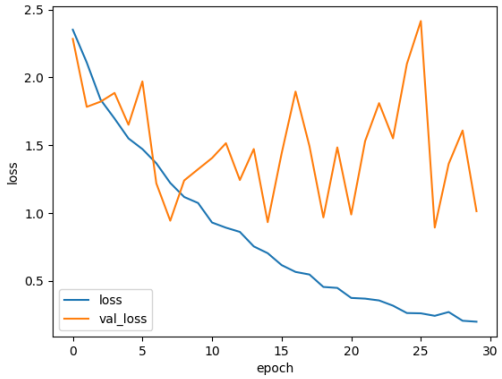


Fig. 5. Loss

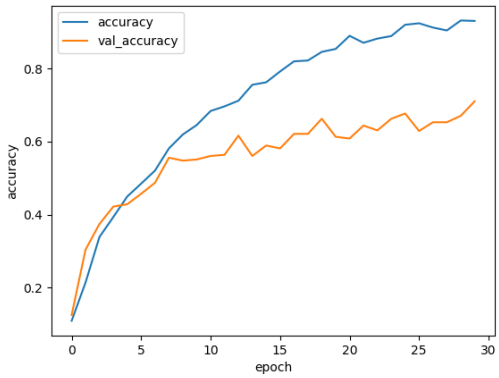


Fig. 6. Accuracy

V. CONCLUSIONS AND FUTURE WORK

We presented a convolutional network for classification of dessert categories and achieve an accuracy of 71.02% for the dessert image dataset. In future, we plan to introduce more dataset and classification scenarios to improve the accuracy of the proposed model.

REFERENCES

[1] Shota Sasano , Xian-Hua Han , Yen-Wei Chen , “Food recognition by combined bags of color features and texture features” , 2016 9th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics.

[2] Weishan Zhang , Dehai Zhao , Wenjuan Gong , Zhongwei Li , Qinghua Lu , Su Yang , “Food Image Recognition with Convolutional Neural Networks , 2015 IEEE 15th Intl Conf on Scalable Computing and Communications and Its Associated Workshops.

[3] Ukrit Tiankaew , Peerapon Chunpongthong , Vacharapat Mettanant , “A Food Photography App with Image Recognition for Thai Food” , 2018 Student Project Conference (ICT-ISPC), ICT International.

[4] Kerry Parker, [online] Available : https://towardsdatascience.com/data-science-skills-web-scraping-using-python-d1a85ef607ed.

[5] Esam Kamal MARIED, Mansour Abdalla ELDALI , Osama Omar ZIADA [online]Available : https://www.researchgate.net/publication/317346042_A_Literature_Study_of_Deep_learning_and_its_application_in_Digital_Image_Processing.

[6] Deepika, Divya , Nayana , Pooja , Dr. Umadevi , “Literature Review of Image Based Online Product Search” , 2015 International Journal of Science, Engineering and Technology Research.

[7] P. Pouladzadeh, G. Villalobos, R. Almaghrabi, S. Shirmohammadi, "A novel svm based food recognition method for calorie measurement applications", ICME Workshops, pp. 495-498, 2012.

[8] Okeke Stephen, Uchenna Joseph Maduh, Sanjar Ibrokhimov, Kueh Lee Hui, Ahmed Abdulhakim AlAbsi, Mangal Sain, “A Multiple-Loss Dual-Output Convolutional Neural Network for Fashion Class Classification”, “International Conference on Advanced Communications Technology(ICACT)” February 2019

[9] Natthawat Phongchit, Available : blog.datawow.io/มาลองดูวิธีการคิดของ-cnn-กัน-e3f5d73eebaa

Extract keywords embedded in comments from YouTube to predict trends

Nattakorn Kointarangkul, Nuttapol Thitaweera, Chittipat Pasomsup
Department of Computer Engineering
Faculty of Engineering
Chulalongkorn University
Bangkok, Thailand
{6270076621, 6270086921, 6270034221}@student.chula.ac.th

***Abstract*— We propose a program written by using TF-IDF algorithm to extract Thai YouTube video keywords. This research is based on identifying Trend Keywords from other keywords by using stop-words algorithm. For the evaluation part, we use various media genres to classify specific types of keywords. As a result, our algorithm was able to identify the targeted keywords and grouped them in accurate genres. The results show that accuracy levels varied between 43% and 69% depend on genre. However, the performance of TF-IDF depends on the numbers of video comments. To sum up, the more keywords we have, the more accurate the prediction will be.**

Keywords*— *TF-IDF; text representation; YouTube; keyword extraction

I. INTRODUCTION

In these modern days, YouTube has become a vital part of our lives. No matter which country you are from, one could not deny the effect it has on us. At this very moment, there are more than 1.9 billion YouTube's Active Users in each month. Furthermore, according to the statistics gathered, YouTube is ranked, "second" in the most popular Search-Engine Category, of which, each user is likely to spend no less than 40 minutes per access on the platform. Interestingly, the time spent on each access for every user is predicted to raise to more than 50% per year. To add on, for each and every poll conducted, Thailand's YouTube users are ranked among the top 10 list of countries, which have highest YouTube users: 93% of the YouTube users in Bangkok and its surrounding counties are active ones, and 92% of the rest of the country are active ones as well. YouTube consequently is considered as one of the strongest marketing channels apart from Facebook. However, it takes a considerable amount of effort to try to follow the YouTube's trends, since they could not be measured precisely and as a result, could not be predicted in an effective manner. Hence, there is still room for improvement, which will make the data from YouTube become more visible and valuable. Our challenge is to bridge that "gap" and come up with a newly invented solution to detect various trends embedded in the platform such as emotions and toxicity. We propose a framework to extract YouTube video keywords in order to capture YouTube's trends from users' comments

II. RELATED WORK

Term frequency-inverse document frequency (TF-IDF) is one of keyword extraction methods [1] that cloud find related words in a document which are embedded in the context. It allows us to get rid of unnecessary words, hence only the targeted ones are left for us to examine further. However, for Thai language, one of the many difficulties is that there exists of no space within and between sentences, which undoubtedly causes the TF-IDF algorithm to be less effective. N. Ousirimaneechai, et al. [2] has proposed an algorithm which would find a stop-words without using the word tokenization or other training data sets and corpora. He presented a solution for finding the Trend Keywords from using Character n-Grams, TF-IDF, K-Means and Elbow method, and validated the result by classifying his output.

Moreover, P. Sarakit et al. [3] considered how to automatically recognize the emotions which were embedded in YouTube videos. They use 6 basic emotions: Anger, Disgust, Fear, Happiness, Sadness, and Surprise. Performances were measured by using 3 alternative machine learning algorithms, namely, multinomial naïve Bayes (MNB), decision tree (DT) and support vector machine (SVM). The results have shown that this particular algorithm could actually classify sentiment and moods in the videos. Moreover, the further study of classifying emotion in YouTube video has been conducted in 2017; J. Savigny, et al. [4] compared the results by weighting with unigram, bigram and trigram methods which involved a number of word embedding techniques such as average word vector, average word vector with TF-IDF, Paragraph vector and Convolutional Neural Network (CNN), of which, illustrated the achievement of extracting the focused contexts in the video comment.

Finally, the social-media contents and their topics are characterized by some keywords and their scale of impact changes over time. As a result of these behaviors, a method to extract and explore the dynamic of data from tweets was proposed. In this paper [5], they proposed a dynamic-keyword extraction method from social media topics, that outperformed all the baselines and produced the value of exploited keywords.

III. METHODOLOGY AND RESULTS

In this work, we propose a framework to Capture YouTube's Trend from Thai YouTube Comments which consist of 7 main steps (A) Collecting YouTube Comment (B) Data Cleansing (C) Word Tokenization (D) Bag of words (BOW) (E) TF-IDF (F) Visualization (G) Evaluation

A. Collecting YouTube Comment

We collect data from YouTube by using YouTube data API V3. API is an Application Programming Interface designed to retrieve various types of data from YouTube such as trendy videos, channels, playlists, and comments

Our targeted data is the comments based within Thai YouTube channels. At this very moment there are a total of 28,500 comments embedded within trending YouTube's video, as shown in Table I.

TABLE I. A SAMPLE OF TRENDING VIDEO COMMENT ON YOUTUBE

Index	Video ID	Category ID	Title	Tags	Comment ID	Text Display
0	xQI9oZEY-B0	10	GOT7 "니가 부르는 나의 이름(You Calling My Name)" M/V	JYP Entertainment|JYP|JB|마크|잭슨|진영|영재|뱀뱀|유겸|GOT...	UgzWu98MUBKTBSmyvIB4AaABAg	1 like = 1 ahgase. ↓ ↓ ↓
28500	mYbyE1cLGl4	10	[เพลงใหม่ล่าสุด 2019] รวมเพลงลูกทุ่งดังฟังสบ...	เพลง|เพลงใหม่|เพลงใหม่ล่าสุด|เพลงลูกทุ่งใหม่ล่...	UgwtG9fJwKjVEUrsc0R4AaABAg	เข้านี่ขอแนะนำ Playlist เพลงเพราะๆ กับ 15 เพลง...

B. Data Cleansing and preparation

In Text classification we use words as one of the main features so it's important to remove unwanted characters such as numbers, emoji, HTML tags. For this research we focus on Thai Language, therefore, none Thai characters will be removed during the process of this step. We used regular expressions for validating Thai Characters, which automatically will help filling in missing comments, as shown in Table II.

TABLE II. RESULT AFTER PROCESSING OF COMMENT

Index	Video ID	Category ID	Title	Tags	Comment ID	Text Display
61	xQI9oZEY-B0	10	GOT7 "니가 부르는 나의 이름(You Calling My Name)" M/V	JYP Entertainment|JYP|JB|마크|잭슨|진영|영재|뱀뱀|유겸|GOT...	Ugzn3adqOE3Ejsw1wpN4AaABAg	ทุกครั้งที่คัมแบ็คพวกคุณไม่เคยทำเรา ผิดหวังเลยแ...
255	xQI9oZEY-B0	10	GOT7 "니가 부르는 나의 이름(You Calling My Name)" M/V	JYP Entertainment|JYP|JB|마크|잭슨|진영|영재|뱀뱀|유겸|GOT...	Ugyt0RfW2sqKz0tG2V14AaABAg	ไทยมารวมกันตรงนี้

C. Word Tokenization

Tokenization is the act of breaking up a sequence of strings into pieces such as words, keywords, phrases, symbols and other elements called tokens. Tokens can be individual words, phrases or even whole sentences. In the process of tokenization, some characters like a punctuation mark, and stop words are discarded we used PyThaiNLP (Thai Natural Language Processing in Python) for word tokenize and remove stop words. An example result is shown in Table III.

TABLE III. RESULT OF WORD TOKENIZATION

Title	Comment ID	Text Display	Tokenized
GOT7 "니가 부르는 나의 이름(You Calling My Name)" M/V	Ugzn3adqOE3Ejsw1wpN4AaABAg	ทุกครั้งที่คัมแบ็คพวกคุณไม่เคยทำเราผิดหวังเลยแ...	[ทุกครั้งที่, คัม, แบ็ค, ไม่เคย, ทำ, ผิดหวัง, ...

D. Bag of words (BOW)

Bag of words (BOW) the bag of words model is a simplified representation used in natural language processing and information retrieval. We will generate our model by applying these following steps. First of all, for each word in a sentence we check if the word exists in our dictionary or not. If it does then we increase the count, on the other hand, if it does not, we will add the word to our dictionary. Secondly, we construct a vector if a word in a sentence a frequent word, we set it as the previous one, else, we set it as 0 but for our research, we decided to use TF-IDF Vector Space Representation.

E. Term Frequency-Inverse Document Frequency

Use the TF-IDF algorithm for keyword extraction on comments made in YouTube: words with high-frequency will be marked as keywords, on the other hand, words with low frequency will be regarded as the opposite. The TF-IDF is an information retrieval technique that weighs a term's frequency (TF) and its inverse document frequency (IDF). TF-IDF is evolved from IDF that proposed by [6] for each word or term has its respective TF and IDF score. The product of the TF and IDF of a term is called the TF-IDF "weight" for The Calculation for TF-IDF that used for term weighting is shown in the equation (1) as provided by [7]

$$(TF\text{-}IDF)_{ij} = TF_{ij} \times Log(IDF_j) \quad (1)$$

Leading to an assumption of their relationship and correlation in one or many ways. Therefore, we could expect an output of emotions and reactions based on a particular video. An example results are shown in Table IV.

TABLE IV. EXAMPLE RESULT OF TF-IDF OF GOT7 "니가 부르는 나의 이름(YOU CALLING MY NAME)" M/V

Keyword	Value
เพลง	0.3103992424729721
ปั่น	0.17970482458961543
วิว	0.1306944178833567
เต้น	0.09802081341251752
ดนตรี	0.04901040670625876

F. Visualization

Mapping a group of words which could be categorized into its genres. Hence a graph theory could be implemented to enable the result to be visualized. To show the relationship between weight and its frequency, the scatter graph enables us to do so by showing the significant keywords below in Fig. 1.

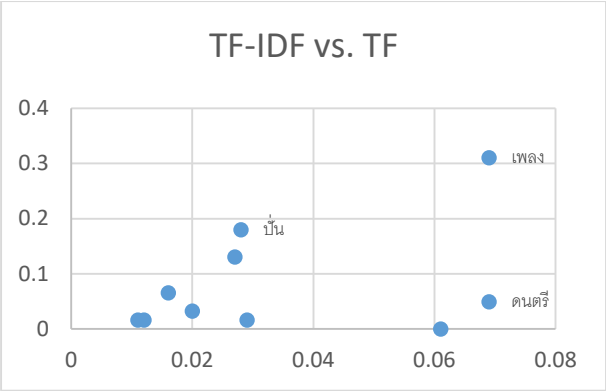


Fig. 1. Relationship between TF-IDF and TF

G. Evaluation

Evaluate the result by experimenting with the "keywords" found and examine their differences. For Evaluation Method we used a human for validating the results of keywords, Fig. 2 show the accuracy of keywords in each genre. Conclusion and future work should be conducted based on previous results.

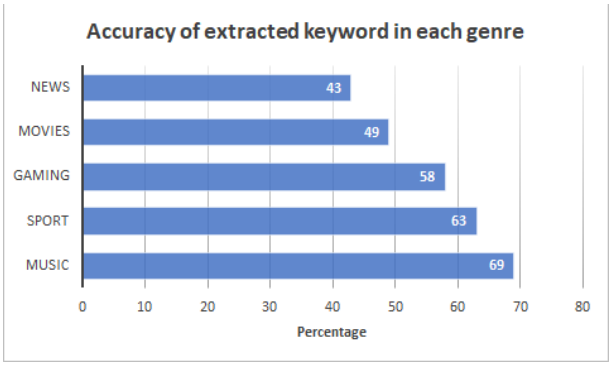


Fig. 2. Accuracy level of keywords in each genre

IV. CONCLUSION AND FUTURE WORK

After implementing the program, we have reached a conclusion that the algorithm has the ability to identify various types of keywords and results show that the lowest accuracy levels is news genre, 43% and the highest is music that is 69%. The reasons that can support this maybe the meaning of words in each genre which not comparable when we separate them into keywords. Therefore, performance of TF-IDF depended on the number of video comments and word tokenizes. In a case where a list of various types of videos are given, the Word Tokenizer performance could significantly be improved, hence the upcoming trend of each video genre could be forecasted more effectively . In the future, we should collect keywords from different video genres, because the pool of keywords may lead to a significant discovery of a new upcoming trend. Moreover, we should be willing to try out other data models to improve the accuracy of the predictions.

ACKNOWLEDGMENT

The authors would like to thank Prof. Prabhas Chongstitvatana for reviewing and editing the content of this paper, and research team for support. This research is sponsored by Department of Computer Engineering, Faculty of Engineering, Chulalongkorn University.

REFERENCES

[1] S. Beliga, A. Meštrović, and S. Martinčić-Ipšić, "An Overview of Graph-Based Keyword Extraction Methods and Approaches," *Journal of Information and Organizational Sciences*, vol. 39, no. 1, pp. 1–20, 2015

[2] N. Ousirimaneechai, and S. Sinthupinyo "Extraction of Trend Keywords and Stop Words from Thai Facebook Pages Using Character n-Grams", *International Journal of Machine Learning and Computing (IJMLC)* Vol. 8 No. 6 pp. 589-594 2018

[3] P. Sarakit, T. Theeramunkong, C. Haruechaiyasak and M. Okumura, "Classifying emotion in Thai YouTube comments," *2015 6th International Conference of Information and Communication Technology for Embedded Systems (IC-ICTES)*, Hua-Hin, 2015, pp. 1-5.

[4] J. Savigny, A. Purwarianti "Emotion Classification on YouTube Comments using Word Embedding", *International Conference on Advanced Informatics, Concepts, Theory, and Applications (ICAICTA)*, 2017

[5] D. Semedo, J. Magalha˜es "Dynamic-Keyword Extraction from Social Media", *European Conference on Information Retrieval (ECIR)* pp 852-860 2019

[6] K. Sparck Jones. IDF term weighting and IR research lessons. *Journal of Documentation*, 60(6), 521-523. 2004.

[7] G. Salton, C. Buckley. " Term-weighting approaches in automatic text retrieval". I*nformation Processing & Management*, 24,(5), pp. 513-523, 1988.

Cashflow prediction of Automated Teller Machine in Bangkok

Thanachai Damrithamnij, Baiboon Permthong, Nutthanit Wiwatbutsiri
Department of Computer Engineering
Faculty of Engineering
Chulalongkorn University
Bangkok, Thailand
sankai1@hotmail.com, baiboon.p@gmail.com, nutthanit.w@gmail.com

***Abstract*—People use ATM service to pay bills, daily spending and withdrawn cash for their personal use. Being able to predict amount of daily cash needed in each ATM is very important to financial institutions as it ensures cash is used efficiently and effectively throughout the branch network. In this paper, we will use the method of moving average to forecast future cash flow from historical data.**

Keywords—ATM, cashflow, moving average

I. INTRODUCTION

ATMs are considered very important tool for interaction between bank and customers. People use ATM service to pay bills, daily spending and withdrawn cash without worrying about bank's working days, business hours, etc. To be able to serve customers when they need cash at any time is beneficial to the bank. However, storing cash in the ATMs have costs, for example, overstocking cash means banks cannot invest this non-earning asset to generate interest income and delivery cost that bank absorb when deliver cash to each ATM. One of the primary concerns with ATM management is to determine the efficient level of cash inventory in each machine and there are different techniques offered to handle this problem.

II. MOTIVATION

ATMs are like major physical laws of open system environments with dynamic cash balancing within the boundaries of the machine. Running out of cash at an ATM or other location means reduced revenue from lost surcharge fees and increased expenses due to emergency currency deliveries. ATM forecasting and services availability is one of the most crucial factors in the ATM network services business. Using ATM cash management optimization and efficient cash loads routing, banks can avoid too much cash in the ATMs. The key to the ATM's forecasting is to capture and process the historical data such that it provides insight into the future.

III. LITERATURE REVIEW

A. ATM Cash Flow Prediction

For many financial institutions, ATMs act as facility that serve their customers for daily needs such as bill payments, transfer of funds, and cash withdrawn. Customers need to be able to withdraw cash for their personal use 24/7. Being able to predict amount of daily cash needed in each ATM is very important to financial institutions as it ensures cash is used efficiently and effectively throughout the branch network. Serengil and Ozpinar (2019) claimed that some banks might store 40% more cash in ATMS than its demand. Finance institutions might have thousands of ATMs. That's why even small optimizations in business operations would contribute high earning. This concern becomes even more critical for countries having high-interest rate and overnight interest rates are higher.

One of the most crucial factors that impact the amount of cash in the ATMs is the location of the ATMs. According to Genevois, Celik, and Ulukan (2015), cash management is one of the main concern of a bank and the problem is to determine how much cash should be put in each ATMs, so it satisfy customer demand. If the demand is higher than the amount of cash in the ATM, then the bank will have to pay for the cost of refilling tasks. Other factors related to the prediction of cashflow in the ATM are population density, population salary, and holiday effect.

B. Cost Function

Running out of cash at the ATM is not an ideal scenario for the bank since it means bank lost revenue in the form of surcharged fee and impact bank reputation. However, by overstocking cash means bank cannot invest this unused cash to generate interest income. Serengil and Ozpinar (2019) claimed that refilling low amount of cash in the ATM would not be a solution because each refilling has a cost of out-of-service time and overtime pay of employees. In their study, they evaluate expected demand, duration, and transportation expenses to find the cost for the following seven days cumulatively. They found both negative interest reflection and transportation cost for a candidate pair of amount and days.

C. Methods

There are many research related to the prediction of cashflow and the cost optimization for refilling cash at the ATM. Bhandari and Gill (2016) used artificial intelligence concept to develop ATM forecasting system. In their study, they used artificial neural networks which provide a methodology for solving many types of non-linear problems that are difficult to solve by traditional techniques. Artificial neural networks can be described as a pool of processing units which communicated among themselves by sending signal. Then each of these units accumulates the input its received, then produce output according to some defined function. The steps they used for ATM forecasting model are as follow. First, they collect past data which contained detail of total amount withdrawn with dates. Then, they normalized such data into values between 0 and 1. The purpose of this step is to allow the activity function to work at least at the beginning of the learning phase. Therefore, the gradient which is a function of the derivative of the nonlinearity will always be different from zero. Next is feature extraction including day

number, week day, weekend effect, salary effect, and holiday effect which also extract from past data.

Next, method used to optimize ATM cash management is generic algorithm. Armenise (2012) suggested the application of generic algorithms as means for searching and generating optimal upload strategies, aimed at identifying a set of uploading rules able to minimize the residual stock and to guarantee service availability at the same time. He believed that generic algorithm is the best search to find the optimal solution. Instead of using generic programming which gives a more general structure, it used higher cost of space than generic algorithm. The data used in this study consists of 30 ATMs which divided into 2 groups of 20 ATMs and 10 ATMs with at least one representative for each class. Also, 70% uptime of ATMs is being recorded that is the time of service availability for cash withdrawals.

Another popular method used for predicting cash flow in ATMs is machine learning. Machine learning code is responsible for predicting daily demand of customers. Serengil and Ozpinar (2019) used neural networks model to predict daily expected cash withdrawals and deposits for the following 15 days. The difficult part is that the model assumes all the workload from the previous n days as inputs. So, today prediction can be done by upload yesterday's workload since we already known the previous day workload. Therefore, tomorrow workload can be identified by catching today's prediction as an input and tomorrow's forecast will be given as an input for the next day prediction and so on.

There are research shows that the location of the ATMs is also very important to predict the cash flow of the ATMs. Chowdhury (2017) studied how we can optimize the location of the ATMs networks. The methodology used is divided into the following sections: data pre-processing, visualization of extracted features, inferring the priority weights to be assigned to each features and deduction. After that we exploit the weights to fit a regression to compute revenues generated by each ATM. The data consisted of 11,229 ATM locations in the state of California and for every zip code 73 more features are observed from the website. For the accuracy of the analysis, they deployed 2 separate models: one for capturing the global features and another for the local features. Both models are used to compute the weight of the features and predictions are computed. They observed that the factors with higher weights vary greatly with the counties.

IV. DATA COLLECTION

Experimentation made use of data collected by a pool of 5 ATMs, chosen to represent a wide range of different location around Bangkok area. The following is the summary of where the ATM are located.

The sample data is collected from a commercial bank based in Thailand. The period of data used is 6 months from January to June 2019. Our sample consist of daily information such as number of customer withdrawn both bank own customers and other bank customers. Also, the amount of money being withdrawn from both bank own customers and other bank customers are collected. Hence, we can see the cash flow movement of each ATM.

TABLE I. ATM ID NUMBER

ID Number	Location
BK1001	BTS Station Saladang
BK1005	Paragon Mall
BK2500	Q-House Office
BK2506	7-11 Store Rama 3
BK3001	Central World Mall

V. PROPOSED MODEL

In this section, we will outline the methodology used to forecast the amount of cash flow needed for each ATM around Bangkok area. Predicting how to fill cash in ATM each day is difficult to do. There are so many factors involved in the prediction – people behavior, weather, rational and irrational behavior of people spending. All these aspects combine to make ATM cash flow prediction very volatile and very difficult to predict with high degree of accuracy. So, we introduce the method of machine learning to help predict the amount of cash flow needed in the next 7 days.

In this research, we introduce the method of "Moving Average" which is a technique often used in technical analysis that smooths price histories by averaging daily prices over some period of time. The predicted cash flow for each day will be the average of a set of previously observed values. Instead of using the simple average, we will be using the moving average technique which uses the latest set of values for each prediction. In other words, for each subsequent step, the predicted values are taken into consideration while removing the oldest observed value from the set. Fig. 1 show a simple illustration that will help understand this with more clarity.

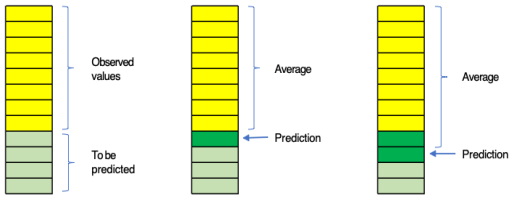


Fig. 1. Moving average framework

Therefore, the first step is to create data frame that only contain only the date and amount of cash withdrawn, then split it into train and validation sets to verify our predictions.

VI. RESULT

The amount of cash withdrawal from ATM depend greatly on what day of the week it is and whether it is the beginning of the holiday period or not. Fig. 2 to 8 below illustrates Paragon Mall withdrawal cashflow each day and the forecasting period (July 2019). On most days, the amount of cash withdrawn increases as holiday period approach. This implied that people tend to withdraw cash to spend during the holiday.

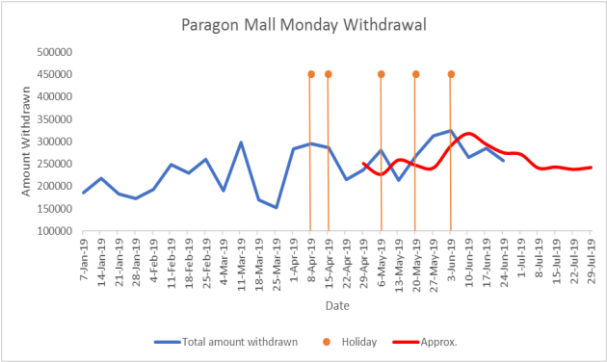


Fig. 2. Paragon Mall Monday Withdrawal

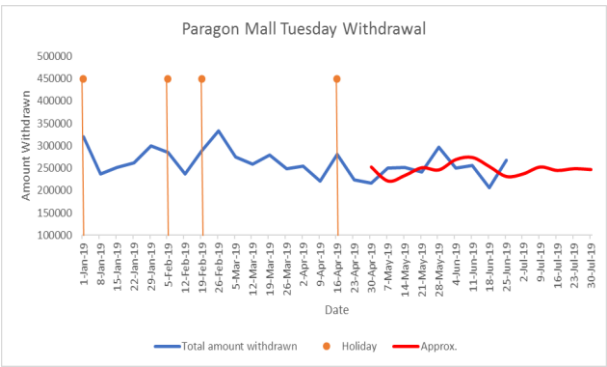


Fig.3. Paragon Mall Tuesday Withdrawal

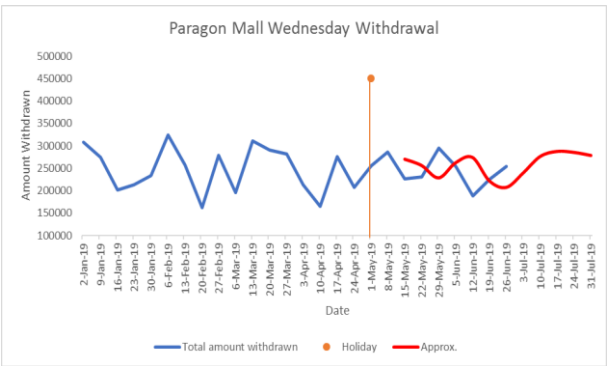


Fig. 4. Paragon Mall Wednesday Withdrawal

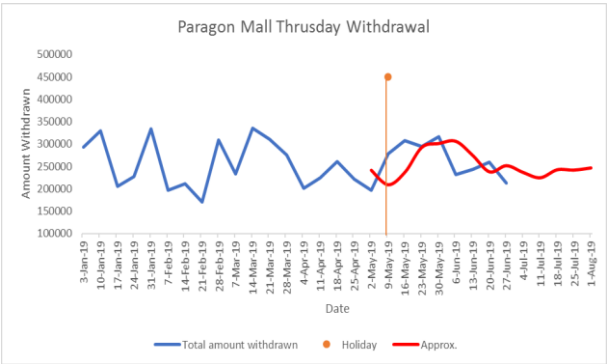


Fig. 5. Paragon Mall Thursday Withdrawal

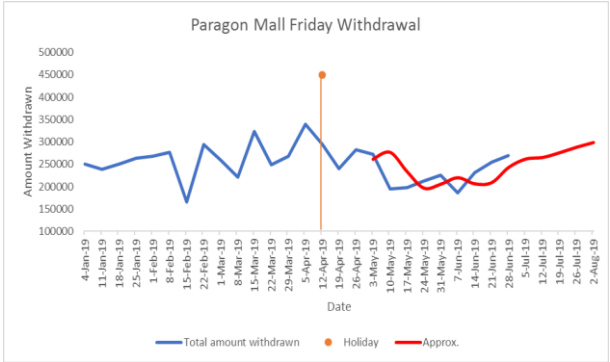


Fig. 6. Paragon Mall Friday Withdrawal

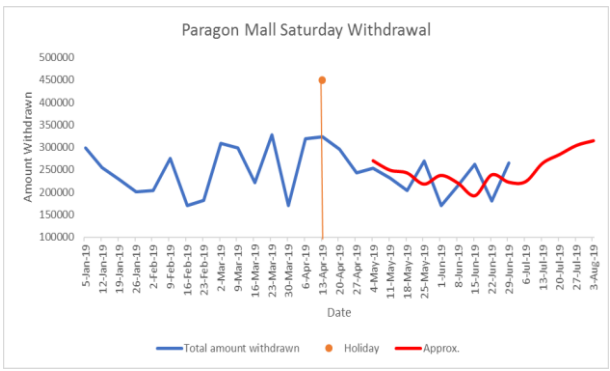


Fig. 7. Paragon Mall Saturday Withdrawal

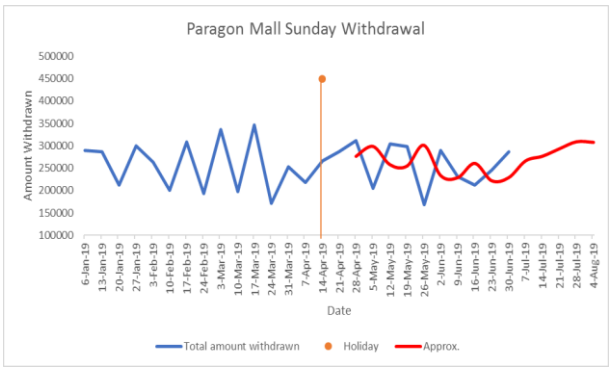


Fig. 8. Paragon Mall Sunday Withdrawal

The result of forecast cashflow in each ATM on each day is summarize in the table below.

TABLE II 7-DAYS CASH FLOW FORECASTING RESULT

		Forecasting (1 - 7 July 2019)						
ATM ID	Location	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
BK1001	BTS Saladang	186,200	170,000	246,400	270,100	212,200	221,400	192,500
BK1005	Paragon Mall	271,400	236,900	239,700	236,300	261,600	285,000	266,600
BK2500	Q-House Office	188,300	220,700	235,300	269,400	282,900	125,000	102,000
BK2506	7-11 Rama 3	216,200	194,500	210,200	184,000	164,800	110,500	98,500
BK3001	Central World	188,000	215,000	220,900	233,700	254,000	271,500	221,000

The forecasting result showed that location and holiday period are important factors to the amount of cash flow in ATM. From our ATM selected for this study, the highest amount of cash been withdrawn is the ATM located at the shopping mall especially on weekend period. As we can see from Fig. 6 and 7, the amount of cash being withdraw on Friday and Saturday

is significant higher than other days. Moreover, during holiday period the amount of cash withdraw increase almost double. On the other hand, the amount of cash is low for the ATM located at the working office during weekend but high on weekdays.

VII. CONCLUSION

To be able to predict ATM cash flow accurately, banks could potentially benefit from it in term of increasing profit in the form of fees or reduce costs such as delivery cost, interest rate cost etc. However, it is a challenging since there are many factors that could influence the amount of cash flow needed in each ATM. Moreover, there are some other factors such as people spending pattern that cannot accurately forecast.

In this paper, we used 6 months historical cash withdrawal data of 5 ATMs around Bangkok area and try to predict next 7 days of cash needed to fill in each ATM. We found that ATM at the shopping mall such as Paragon and Central World mall need the highest amount of cash on the weekend. As for ATM at the public transport such as BTS station, we need to fill the cash during weekdays and not much on weekends. As for ATM at the working office area, filling cash before holiday period is recommended.

In this study, we use the method of moving average for forecasting future cash flow. There are drawbacks to this methodology. Since moving average draw trends from past information, they do not consider changes that may affect the amount of cash being withdraw such as new competitor ATM being place around that area. Moreover, moving average does not consider any seasonal change that past data represented. Therefore, for future study one might try to use time series approach to forecast amount of cash needed in ATM.

REFERENCES

[1] Sefik Ilkin Serengil, Alper Ozpinar (2019). ATM Cash Flow Prediction and Replenishment Optimization with ANN. International Journal of Engineering Research and Development. Volume 11

[2] Somnath Basu Roy Chowdhury, Biswarup BhaŠacharya, Sumit Agarwal (2017). Location Optimization of ATM Networks.

[3] Renu Bhandari, Jasmeen Gill (2006). An Artificial Intelligence ATM forecasting system for Hybrid Neural Networks. International Journal of Computer Applications (0975 – 8887). Volume 133 – No.3.

[4] M. Erol Genevois, D. Celik, H. Z. Ulukan (2015). ATM Location Problem and Cash Management in Automated Teller Machines. World Academy of Science, ngineering and Technology International Journal of Industrial and Manufacturing Engineering. Volume 9, No:7.

[5] Roberto Armenise, Cosimo Birtolo, Eugenio Sangianantoni, and Luigi Troiano (2012). Optimizing ATM Cash Management by Genetic Algorithms. International Journal of Computer Information Systems and Industrial Management Applications. Volume 4.

Violence detection from surveillance camera

Nun Vanichkul, Supachai Jiamwijitkul, Chonbadee Juthamanee
Department of Computer Engineering
Faculty of Engineering
Chulalongkorn University
Bangkok, Thailand
{Nun.Vanichkul, Supachai.Jiamwijitkul, Chonbadee.Juthamanee}@chula.ac.th

บทคัดย่อ **งานวิจัยนี้นำเสนอ ระบบตรวจจับเหตุการณ์รุนแรงจากกล้องวงจรปิด เพื่อเพิ่มความรวดเร็วและความเป็นอัตโนมัติในการแจ้งเหตุการณ์ ซึ่งได้ใช้การวิเคราะห์และพัฒนาวิธีการมากจากเทคนิค Bag of Visual Word เพื่อเพิ่มความแม่นยำให้มากขึ้น โดยนำข้อมูลคลิปวิดีโอจำนวน 1,000 คลิปที่มีความรุนแรง และอีก 1,000 คลิปที่ไม่มีความรุนแรงมาใช้ในการสร้าง model ซึ่งให้ค่าความแม่นยำในการตรวจจับความรุนแรงอยู่ที่ 90% รวมถึงวิเคราะห์ข้อจำกัดและความท้าทายจากผลการทดลอง**

***Abstract*—This work proposes a system to detect violence events from surveillance camera in order to improve the speed and autonously report the event. The analysis is based on the technique of bag of visual words. The dataset consists of 1,000 video clips that contains violence scenes and 1,000 video clips that does not contain violence scenes. The accuracy of the model is 90%. The limit and the challenge of the work is discussed.**

Keywords—Video surveillance, image processing, security

I. INTRODUCTION

ปัจจุบันเจ้าหน้าที่รักษาความปลอดภัยที่คอยเฝ้าสังเกตภาพจากกล้องวงจรปิดไม่สามารถจดจ่อเฝ้าสังเกตได้อย่างต่อเนื่องเป็นเวลานานจากอาการล้าของหรือช่วงที่ละเลยการปฏิบัติหน้าที่อย่างขะมักเขม้น ซึ่งเป็นส่วนหนึ่งของสาเหตุให้การติดตามและระงับเหตุฉุกเฉินเกิดข้อบกพร่อง ดังนั้นเพื่อให้กล้องวงจรปิดมีความสามารถในการสังเกตรูปที่ถ่ายและทำการวิเคราะห์จนสามารถระบุได้ว่าเหตุการณ์ที่กล้องจับภาพอยู่นั้นเป็นเหตุการณ์ผิดปกติสมควรจะส่งสัญญาณแจ้งให้กับทางเจ้าหน้าที่ทราบเพื่อเฝ้าระวังได้อย่างทันท่วงที งานวิจัยนี้จึงได้นำเสนอ ระบบตรวจจับเหตุการณ์รุนแรงจากกล้องวงจรปิด เพื่อเพิ่มความรวดเร็วในการแจ้งเหตุการณ์

II. REALATED WORK

ระบบการเฝ้าระวังด้วยภาพวิดีโออัตโนมัติเป็นหัวข้อที่ได้รับความสนใจเป็นอย่างมากในงานวิจัยที่เกี่ยวกับการรับรู้ภาพของคอมพิวเตอร์ซึ่ง T. Ko [1] ได้ทำการสำรวจงานวิจัยต่าง ๆ ที่เกี่ยวข้อง และได้ทำการสรุปกรอบการทำงาน รวมถึงจัดกลุ่มให้แก่วิธีการซึ่งเป็นที่นิยมในส่วนงานแต่ละส่วนไว้ดังนี้

A. การตรวจจับการเคลื่อนไหวและวัตถุ

เป็นการระบุถึง pixel ซึ่งบ่งบอกการเคลื่อนที่ของวัตถุที่สนใจในภาพนิ่งหรือ เฟรมแต่ละเฟรมซึ่งเป็นหน่วยย่อยของภาพเคลื่อนไหวที่กล้องวีดีโอถ่ายไว้ได้ โดยมีวิธีการที่นิยมได้แก่ การเปรียบเทียบความแตกต่างระหว่างแต่ละ pixel ในแต่ละเฟรมที่มีลำดับต่อเนื่องกัน (Temporal differencing) การเทียบความแตกต่างระหว่างภาพพื้นหลังซึ่งบันทึกไว้ก่อนหน้า กับรูปที่ถ่าย ณ ขณะใด ๆ (Background subtraction) และ Optical flow

B. การจำแนกวัตถุ

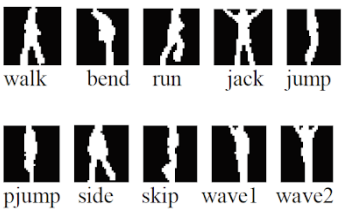
สามารถมองเป็นโจทย์การหารูปแบบของวัตถุต่าง ๆ เพื่อจำแนกเฉพาะวัตถุที่สนใจได้ โดยวิธีการที่นิยมสามารถแบ่งได้ 2 วิธี ได้แก่ การจำแนกจากรูปลักษณ์ (shape-based classification) และ การจำแนกจากการเคลื่อนไหว (motion-based classification) โดยวิธีการหนึ่งในการจำแนกมนุษย์ ออกจากภาพพื้นหลังได้แก่ Histograms of Oriented Gradient (HOG) ควบคู่กับการใช้ linear Support Vector Machine [2] ซึ่งมีความแม่นยำประมาณ 89% ในการบ่งชี้ว่ามีมนุษย์อยู่ในภาพขนาด 64 X 128 pixel หรือไม่

C. การติดตามวัตถุ

เป็นการหาความสัมพันธ์ระหว่างวัตถุที่ปรากฏในเฟรมหนึ่งกับวัตถุที่ปรากฏในอีกเฟรมหนึ่งซึ่งจำแนกวิธีการได้เป็น 4 กลุ่มใหญ่คือ การใช้กลุ่ม pixel ทั้งหมดของวัตถุ การใช้เฉพาะเส้นแสดงรูปร่างของวัตถุ การใช้คุณลักษณะของวัตถุ และ การใช้แบบจำลองในการติดตาม

D. การสกัดคุณลักษณะของข้อมูลการเคลื่อนไหว

เพื่อให้สามารถรับรู้การเคลื่อนไหวของวัตถุได้ จำเป็นต้องสกัดเอาข้อมูลของการเคลื่อนไหวจากภาพให้ได้ก่อนเป็นอย่างแรก โดยข้อมูลหรือคุณลักษณะที่นิยมใช้ได้แก่ คุณลักษณะการไหลของแสง (Optical Flow Features) คุณลักษณะของเส้นโคจร (Trajectory-based Features) และคุณลักษณะของพื้นที่หรือภาพ (Region-or Image-based Features) โดยตัวอย่างวิธีการสกัดคุณลักษณะการเคลื่อนไหวเช่น วิธี Self-Organizing Map (SOM) เป็นการใช้ Unsupervised Learning Neural Network มาประยุกต์ใช้กับ Image Analysis และ Pattern Recognition อีกวิธีหนึ่งคือ Sequences of Human Silhouettes เป็นการใช้ลักษณะของรูปร่างของมนุษย์ในรูปแบบเงา (Human Silhouettes) โดยแปลงลักษณะของเงาในรูปแบบ 2 มิติ (2D) ให้อยู่ในรูปของ Vector เพื่อแปรผลลักษณะของท่าทางต่าง ๆ ด้วยการหาทิศทางการเคลื่อนที่ของ vector ในรูปแบบ (Action recognition using longest common subsequence scheme) ดังรูปที่ 1



รูปที่ 1 ลักษณะท่าทางของมนุษย์รูปแบบเงา

E. การวิเคราะห์และเข้าใจพฤติกรรม

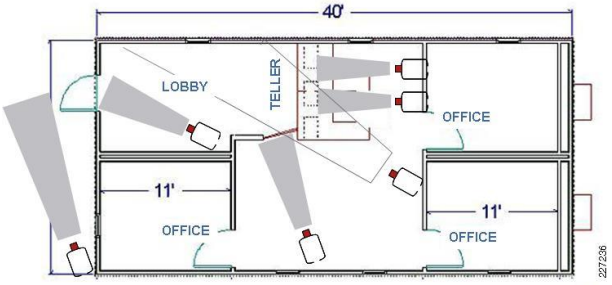
ถือได้ว่าเป็นงานที่ท้าทายที่สุดอย่างหนึ่งในระบบทั้งหมด โดยธรรมชาตินั้นมนุษย์สามารถเข้าใจถึงอารมณ์ของกันและกันได้ด้วยสัญชาติญาณ แต่การเข้าใจพฤติกรรมจากข้อมูลภาพนั้นต้องมีการกำหนดส่วนของร่างกายที่แสดงถึงอารมณ์ต่าง ๆ อย่างแม่นยำ เพื่อให้สามารถเข้าใจสารที่ร่างกายมนุษย์สื่อออกมาเป็นท่าทาง และต้องมีการนำข้อมูลท่าทางเหล่านี้มาวิเคราะห์เพื่อหารูปแบบและวิธีในการแปลความหมายอย่างถูกต้อง ซึ่งในปัจจุบันได้มีงานวิจัยที่เสนอวิธีการต่าง ๆ มากมาย ได้แก่ Hidden Markov Models, Dynamic Time Warping, Finite-State Machine, Nondeterministic-Finite-State Automation, Time-Delay Neural Network, Syntactic /Grammatical Techniques, Self-Organizing Neural Network, Agent-Based Techniques และ Artificial Immune Systems

F. การระบุตัวตน

งานวิจัยส่วนใหญ่นั้นจะทำการระบุตัวตนด้วยการวิเคราะห์และเปรียบเทียบเอกลักษณ์ของแต่ละบุคคลในภาพด้วยข้อมูลต่าง ๆ เช่น การเคลื่อนไหว ลักษณะท่าทาง สัดส่วนของร่างกาย และความเร็วในการหมุนของข้อต่อ เป็นต้น

G. การตั้งกล้องและรวมข้อมูลจากกล้องหลายตัว

ระบบกล้องติดตาม (Surveillance Camera System) กล้องติดตามเหตุการณ์เป็นระบบสำหรับสนับสนุนการเพิ่มความปลอดภัยให้แก่ชีวิต และทรัพย์สิน ซึ่งอาจเป็นของส่วนตัว หรือเป็นบริการจากส่วนกลางหรือภาครัฐก็ได้ ซึ่งอุปกรณ์เหล่านี้เป็นส่วนสำคัญในการได้มาของข้อมูลในหลากหลายมิติ และเป็นส่วนสำคัญในการคาดการณ์เพื่อป้องกันหรือระงับเหตุได้อย่างทันท่วงที โดยในการติดตามเหตุการณ์ จำเป็นต้องกำหนดจุดให้ครอบคลุมพื้นที่ ณ บริเวณหนึ่ง เพื่อเพิ่มจำนวนการได้มาของข้อมูลให้ครบถ้วน หากเกิดกรณีเหตุการณ์อาชญากรรมในจุดที่กล้องตัวหนึ่งไม่สามารถติดตามได้ จะส่งผลให้การระงับเหตุไม่เป็นผล



รูปที่ 2 ตำแหน่งการติดตั้งกล้องให้ครอบคลุมพื้นที่

การรับข้อมูลจากกล้องในระบบกล้องติดตาม จะประกอบไปด้วย 3 ส่วนหลัก ได้แก่

- กล้องติดตาม ใช้สำหรับการติดตามสถานการณ์ และเก็บข้อมูล (Input)
- จอมอนิเตอร์ ใช้สำหรับการติดตามสถานการณ์ด้วยมนุษย์ และแสดงผลข้อมูลการแจ้งเตือนเหตุ
- ระบบเซิฟเวอร์และฐานข้อมูล สำหรับการนำเข้าและประมวลข้อมูล

โดยกล้องวงจรปิดเชื่อมต่อ Wi-Fi กับ router และ นำข้อมูลที่ได้ไปเก็บและประมวลที่ข้อมูลที่ server ซึ่งเชื่อมต่อกับ router ผ่าน Lan หลังจากนั้นก็นำข้อมูลที่ได้ไปแสดงผลการติดตามสถานการณ์และข้อมูลการแจ้งเตือนที่หน้าจอมอนิเตอร์ซึ่งเชื่อมต่อกับ router ผ่าน Lan ตามรูปที่ 3

รูปที่ 3 ระบบกล้องติดตาม (Surveillance Camera System)

H. การวิเคราะห์ประสิทธิภาพ

ในขั้นตอนต่าง ๆ ของระบบนั้นจะนิยมใช้ชุดข้อมูล Performance Evaluation for Tracking and Surveillance ซึ่งรวบรวมภาพถ่ายของทั้งมนุษย์ ยานพาหนะ ซึ่งบันทึกทั้งด้วยกล้องเดี่ยว และกล้องหลายตัว ในสภาพแวดล้อมทั้งในร่มและกลางแจ้ง รวมไปถึงข้อมูลการจำแนกท่าทางของมือ การแสดงทางสีหน้า และการมอง นอกจากนั้นยังมีข้อมูลรูปที่บันทึกการตั้งใจขโมยของ ซึ่งบันทึกจากกล้องหลายตัวในมุมมองที่แตกต่างกัน และการทำตัวมีพิรุธเพื่อหาโอกาสในการขโมยของอีกด้วย

เพื่อเพิ่มประสิทธิภาพในการจำแนกมนุษย์ในภาพจากกล้องซึ่งพื้นที่ส่วนใหญ่จะเป็นพื้นหลังและภาพของมนุษย์มีขนาดค่อนข้างเล็กเทียบกับขนาดภาพทั้งหมด ได้มีการเสนอแนวคิดในการหาจุดเด่นที่คาดว่าจะเป็นภาพของมนุษย์ในเบื้องต้นด้วยเทคนิค Deep Multi-Level Network [3] ก่อน แล้วจึงนำส่วนของรูปที่เป็นจุดเด่นนั้นไปจำแนกอย่างละเอียดว่ามีมนุษย์หรือไม่ด้วยวิธีการ HOG ซึ่งเป็นการลดภาระการคำนวณลงอย่างมาก ทำให้สามารถจำแนกมนุษย์ในภาพได้ด้วยระยะเวลาที่สั้นลง เพิ่มความเป็นไปได้ที่ระบบจะจำแนกมนุษย์ได้อย่างทันท่วงทีในการใช้งานจริง

III. DATASET

Dataset ในการทดลองนี้จะใช้ข้อมูลจาก Kaggle โดยเป็นคลิปวีดีโอจากเว็บไซต์ YouTube [9] ซึ่งแบ่งเป็นคลิปที่ถูกจำแนกว่าบันทึกเหตุการณ์ความรุนแรงจำนวน 1,000 คลิป และคลิปอีก 1,000 คลิปซึ่งไม่มีเนื้อหาเกี่ยวกับเหตุการณ์รุนแรง

ตัวอย่างภาพจากคลิปวีดีโอในรูปที่ 4 ประกอบด้วย 3 รูปด้านบนมาจากคลิปที่มีความรุนแรง และ 3 รูปด้านล่างมาจากคลิปซึ่งไม่มีความรุนแรง

รูปที่ 4 ตัวอย่างภาพจากคลิปวีดีโอ

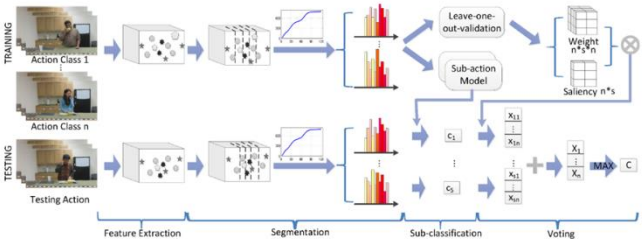
จากข้อมูลที่มีความสมดุลกันทั้งสองจำพวก จะทำให้สร้างโมเดลที่ไม่ลำเอียงในการจำแนกภาพได้ โดยจากข้อมูลจำนวน 2,000 คลิป จะทำการแบ่งออกเป็น 3 กลุ่มด้วยกัน ได้แก่ train data, validate data และ test data ซึ่งแต่ละกลุ่มจะมีจำนวนคลิปเป็นร้อยละ 60, 20 และ 20 ตามลำดับ โดยในแต่ละกลุ่มจะมีจำนวนคลิปวีดีโอที่มีความรุนแรงและไม่มีความรุนแรงเท่ากัน

IV. ANALYSIS

การทดลองนี้จะทำการจำแนกคลิปที่มีความรุนแรงออกจากคลิปที่ไม่มีความรุนแรง โดยจากการสังเกตข้อมูลในคลิปที่มีอยู่ พบว่าในเหตุการณ์ที่เกิดการทำร้ายร่างกายหรือการต่อสู้กันนั้น จะเป็นการที่คนอย่างน้อย 2 คน มีการชกต่อย หรือปลุกปล้ำกัน โดยแขนหรือขาของคนที่ทำการต่อสู้มักจะมีการขยับเคลื่อนไหวอย่างรวดเร็ว เช่นการต่อย การผลัก การตบ การจับทุ่ม ซึ่งจะแตกต่างจากเหตุการณ์ที่ไม่มีความรุนแรงที่คนสองคนอาจจะอยู่ใกล้กัน เช่น การกอดเป็นต้น ซึ่งจากข้อสังเกตดังกล่าวจึงได้เป็นหลักการเบื้องต้นที่อาจใช้บ่งชี้ว่ามีการต่อสู้หรือไม่ คือ 1. คนอย่างน้อย 2 สองคนต้องอยู่ใกล้กันในระยะเอื้อมถึงจึงจะทำการต่อสู้กันได้ 2. มือหรือเท้าของผู้ร่วมเหตุการณ์รุนแรงมักมีการเคลื่อนไหวอย่างรวดเร็ว 3. การที่แขนหรือขาของคนใดคนหนึ่งเหวี่ยงใส่อีกคนหลายครั้ง เป็นการบ่งบอกถึงความตั้งใจที่จะต่อยหรือเตะอีกคนหนึ่ง

จากข้อสังเกตทั้ง 3 ข้อ จึงได้นำมาเป็นตั้งเป็นสมมติฐานที่ว่า หากสามารถสกัดคุณลักษณะจากคลิปวีดีโอซึ่งบ่งบอกได้ว่า มีอย่างน้อย 2 คน อยู่ใกล้กันในระยะเอื้อมถึง ซึ่งบุคคลดังกล่าวมีการเคลื่อนไหวของแขนหรือขาอย่างรวดเร็ว อีกทั้งเป็นการเคลื่อนไหวในทิศทางที่ไปกระทบอีกคนหนึ่งซ้ำกันหลายครั้ง น่าจะช่วยให้การจำแนกเหตุการณ์รุนแรงซึ่งเป็นการต่อสู้กันของคนอย่างน้อย 2 คนได้แม่นยำมากขึ้น

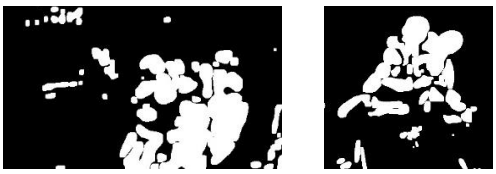
จากสมมติฐาน นำมาสู่วิธีการที่จะนำเสนอ โดย E. Bermejo et al [6] ได้นำเอาวิธีการ Bag of Words ซึ่งประยุกต์มาจากเทคนิคในทางภาษาศาสตร์มาใช้ โดยขั้นตอนแรกจะใช้วิธีการ Motion SIFT (MoSIFT) [7] และ Space-Time Interest Point (STIP) [8] ในการสกัด visual word หรือก็คือการมองคุณสมบัติของพื้นที่ที่สนใจในภาพจากคลิปวีดีโอเป็นเสมือนคำศัพท์ และนำคำ หรือ visual word เหล่านั้นทุกๆ คำจากแต่ละคลิปวีดีโอมาทำการ clustering ด้วย K-mean เพื่อหา centroid ของแต่ละ cluster แล้วจึงสร้าง histogram ของแต่ละคลิปวีดีโอด้วยความถี่ของ visual word ที่ถูกจำแนกเป็นแต่ละ cluster และสุดท้าย histogram ของแต่ละคลิปจะถูกนำมาแบ่งกลุ่มด้วยเทคนิค Support Vector Machine (SVM) เพื่อบ่งชี้ว่าเป็นคลิปที่มีความรุนแรงหรือไม่ โดยจากวิธีการดังกล่าวจะเป็นการนำวีดีโอทั้งหมดมาเข้ากระบวนการสกัดคุณลักษณะ โดยไม่ได้คำนึงถึงข้อสมมติฐานที่กล่าวมาข้างต้นเลย ดังนั้นการทดลองนี้จึงเสนอให้มีการสกัดคุณลักษณะของ การมีคน 2 คนซึ่งอยู่ใกล้กัน แขนหรือขามีการเคลื่อนไหวอย่างรวดเร็วในทิศทางที่ไปกระทบอีกคนหนึ่ง ออกมาเป็นหนึ่งใน visual word ของแต่ละคลิปวีดีโอ แล้วจึงค่อยดำเนินการตามขั้นตอนของวิธีการ Bag of Words ต่อไป



รูปที่ 5 ขั้นตอนในการจำแนกรูปภาพด้วยวิธีการ Bag of Wards [10].

V. EXPERIMENTAL RESULT

ผลการวิจัยนี้สามารถจำแนกคลิปที่มีความรุนแรงออกจากคลิปที่ไม่มีความรุนแรงได้ถูกต้อง 90% ภาพด้านล่างจะแสดงตัวอย่างคลิปวีดีโอที่ถูกจำแนกว่ามีความรุนแรง ซึ่งทั้งสองภาพมาจากวีดีโอที่มีคนโดนล็อคคอจากด้านหลัง



รูปที่ 6 ตัวอย่างคลิปวีดีโอที่จำแนกว่ามีความรุนแรง

ผลการทดลองจากวิจัยชิ้นนี้พบว่าประเภทของคลิปที่ไม่สามารถแยกได้ คือ คลิปประเภทที่มืดเกินไปหรือมีแสงมากเกินไป เพราะทำให้การ detect มีความคลาดเคลื่อน

REFERENCES

[1] T. Ko, “A survey on Behavior Analysis in Video Surveillance for Homeland Security Applications,” in 2008 37th IEEE Applied Image Pattern Recognition Workshop, Washington DC, USA, 2008.

[2] N. Dalal and B. Triggs, “Histograms of Oriented Gradients for Human Detection,” in 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05), San Diego, CA, USA, 2005.

[3] V. Gajjar, A. Gurnani and Y. Khandhediya, “Human Detection and Tracking for Video Surveillance: A Cognitive Science Approach,” in 2017 IEEE International Conference on Computer Vision Workshops (ICCVW), Venice, Italy, 2017.

[4] S. Yonemoto, D. Arita and R. Taniguchi, “Real-time Human Motion Analysis and IK-based Human figure control,” in Proceedings Workshop on Human Motion, Austin, Texas, USA, 2000.

[5] W. Huang and J. Wu, “HUMAN ACTION RECOGNITION BASED ON SELF ORGANIZING MAP,” in 2010 IEEE International Conference on Acoustics, Speech and Signal Processing, Dallas, TX, USA, 2010.

[6] E. Bermejo, O. Deniz, G. Bueno and R. Sukthankar, “Violence Detection in Video Using Computer Vision Techniques”, in Real

P., Diaz-Pernil D., Molina-Abril H., Berciano A., Kropatsch W. (eds) Computer Analysis of images and Patterns. CAIP 2011. Lecture Notes in Computer Science, vol 6855. Springer, Berlin, Hiedelberg, 2011.

[7] M. Chen and A. Hauptmann, "MoSIFT: Recognizing Human Actions in Surveillance Videos," CMU-CS-09-161., Carnegie Mellon University, Pittsburgh, USA, 2009.

[8] I. Laptev, "On Space-Time Interest Points," in Proceedings Ninth IEEE International Conference on Computer Vision, Nice, France, 2003.

[9] Accessed at https://www.kaggle.com/mohamedmustafa/real-life-violence-situations-dataset.

[10] H. Liu, H. Tang, W. Xiao, Z. Guo, L. Tian and Y. Gao, "Sequential Bag-of-Words model for human action classification," in CAAI Transaction on Intelligence Technology, vol. 1, Issue 2, pp. 125-136, Pecking University, Shenzhen Graduate School, China, 2016

YouTube video classification of violent content based on comments

Grace Panitchakorn, Tossapon Nuanchuay
Department of Computer Engineering
Faculty of Engineering
Chulalongkorn University
Bangkok, Thailand
grace.phanitchakorn@gmail.com, tossapon.nuanchuay@agoda.com

***บทคัดย่อ*—ปัจจุบัน YouTube สามารถการคัดกรองเนื้อหาที่ไม่เหมาะสมได้ โดยการตั้งค่าโหมดจำกัดหรือ Restrict Mode โดยดูจากวีดีโอ ชื่อเรื่อง คำอธิบาย และภาษาที่ใช้ ซึ่งวิจัยนี้นำเสนอการคัดกรองเนื้อหาความรุนแรงจากวีดีโอ YouTube โดยใช้ไม่ใช้ข้อมูลวีดีโอ แต่ใช้ข้อมูลการแสดงความคิดเห็นในหน้าเพจจำแนกว่าเนื้อหานั้นมีความรุนแรงหรือไม่มีความรุนแรง ด้วย Convolutional Neural Network การทดลองพบว่าความแม่นยำสูงถึง 84%**

***Abstract* -- YouTube viewers can limit the view of inappropriate content with the restrict mode using title, description and language. This work proposes a method to filter the content without inspection on video data but using comments of that video to classify violence/not-violence. The technique employs Convolutional Neural Network. The accuracy achieved is 84%.**

Keywords— YouTube, text processing, Convolutional Neural Network

I. INTRODUCTION

ปัจจุบันสื่อออนไลน์ได้เข้ามามีบทบาทเป็นอย่างมากในสังคมไม่ว่าจะเป็นแหล่งและเปลี่ยนความคิดเห็น แบ่งปันความรู้ เผยแพร่ข่าวสาร ตลอดไปจนถึงบันทึกเรื่องราวส่วนตัว ถูกนำเสนอเป็นข้อความ รูปภาพ หรือวีดีโอขึ้นอยู่กับประเภทของสื่อออนไลน์นั้น เช่น Facebook, Twitter, Instagram และ YouTube

การใช้งานสื่อเหล่านี้ก่อให้เกิดข้อดีและข้อเสียมากมายทั้งในแง่ของความรวดเร็วในการสืบค้นข้อมูล การเข้าถึงข้อมูลในแหล่งที่กว้างขึ้นเป็นสากลอย่างไร้ขีดจำกัด แต่ข้อดีเหล่านั้นยังเป็นข้อเสียในเวลาเดียวกันหากใช้ในทางที่ไม่ถูกต้องและผู้ใช้ออนไลน์ทั้งผู้ส่งสารและผู้รับสารขาดการวิเคราะห์แยกแยะสื่อสร้างสรรค์และสื่อทางลบ ทั้งนี้การรับรู้ถึงความมีประโยชน์และความง่ายในการใช้งานตัวอย่างกรณีศึกษา YouTube จากการวิจัยที่ศึกษาทัศนคติและพฤติกรรมการสื่อสารผ่านเครือข่ายสังคมในการสร้างชื่อเสียงต่อบุคคลพบว่า ผู้คนในสังคมจะให้ความสำคัญต่อการสื่อสารผ่าน YouTube เพิ่มสูงขึ้น ระยะเวลาและลักษณะการใช้งานรวมถึงกิจกรรมที่ใช้และการใช้งานในอดีตส่งผลต่อการตัดสินใจกระทำการในอนาคตอีกด้วย Araya Giatgong[6]

งานวิจัยนำเสนอกระบวนการจำแนกความความรุนแรงของเนื้อหาของวีดีโอ YouTube จากข้อมูลประกอบวีดีโอบน คือ ความคิดเห็นจากผู้ชมโดยปราศจากการพิจารณาเนื้อหาของวีดีโอโดยตรงเพื่อเป็นทางเลือกในการบ่งชี้ความรุนแรงของเนื้อหาในวีดีโอ เนื่องจากเว็บไซต์ YouTube เป็นสื่อออนไลน์ที่เป็นที่นิยม สามารถเผยแพร่ข้อมูลและรับรู้ได้ง่าย อีกทั้งข้อมูลจากการแสดงความเห็นนั้นจะต้องมาจากบัญชีผู้ใช้ที่ลงทะเบียนแล้วเท่านั้นทำให้ข้อมูลมีการคัดกรองจากนโยบายของเว็บไซต์ในเบื้องต้นแล้ว

งานวิจัยพิจารณาและให้ความสำคัญถึงความแม่นยำในการวิเคราะห์เพื่อให้การป้องกันและการแก้ไขสถานะการจากการเผยแพร่วีดีโอที่มีความรุนแรง ตัวอย่างเช่น การป้องกันการเผยแพร่วีดีโอเหตุกราดยิง วีดีโอเหตุการณ์ในเหตุก่อการร้ายจากมุมมองของกลุ่มก่อการร้าย เป็นต้น

II. LITERATURE REVIEW

อัลกอริทึมและกระบวนการที่สามารถนำมาใช้ในงานวิจัยมีด้วยกันหลากหลายวิธีและสามารถแย่งออกเป็นหมวดหมู่ คือ การดึงข้อมูล, การคัดกรองข้อมูล, การให้คะแนนความรู้สึก และการเรียนรู้ของเครื่องเพื่อบ่งชี้

การดึงข้อมูลจาก YouTube สามารถทำผ่านการเรียกใช้ API ที่ YouTube เป็นผู้กำหนด[1] และมีชุดคำสั่ง 3rd party ให้ใช้มากมาย

การคัดกรองข้อมูล[1] แบ่งออกเป็น 4 ลำดับขั้นตอนคือ การตัดคำ, การลบคำหยุดภาษาไทยและภาษาอังกฤษ, การลบชุดคำสั่ง HTML และการตัดคำ โดยการตัดคำสามารถทำได้โดยการใช้โปรแกรมประยุกต์ LexTo (Thai Lexeme Tokenizer) และพจนานุกรม Lexitron ที่ให้บริการโดย NECTEC

การให้คะแนนความรู้สึกสามารถทำได้จากชุดข้อมูล SentiWordNet [2] เป็นคลังคำศัพท์บอกความรู้สึกที่นิยมใช้ในภาษาอังกฤษและใช้ Wisesight-sentiment เป็นคลังคำศัพท์บอกความรู้สึกภาษาไทย

เทคนิคการเรียนรู้ของเครื่องที่ใช้ในงานวิจัยประกอบไปด้วย Multinomial naïve Bayes (MNB)[1][4][5] , Decision tree (DT), Support Vector Machine (SVM)[1][4] , K-Nearest Neighbor (KNN)[4][5] และ ID3[5]

III. METHODOLOGY

A. เก็บรวบรวมชุดตัวอย่างวีดีโอ

เก็บที่อยู่เว็บไซต์ตัวอย่างของวีดีโอที่มีความรุนแรง และไม่มีความรุนแรง จำนวน 33 วีดีโอ และแบ่งแยกความรุนแรงเนื้อหาจากการตัดสินใจของมนุษย์โดยอ้างอิงถึงนโยบายเกี่ยวกับเนื้อหาที่มีความรุนแรงหรือสยดสยองจาก YouTube[12] และทำการรวบรวมข้อมูลการแสดงความคิดเห็นในหน้าเว็บผ่าน YouTube Data API[9]

B. เตรียมข้อมูลประเภทข้อความ

ขั้นตอนในการประมวลผลข้อความที่ปรากฏในที่อยู่วีดีโอ โดยทำการรวบรวมความคิดเห็น โดยใช้เทคนิคการประมวลผลภาษาธรรมชาติ (Natural Language Processing - NLP) เนื่องจากข้อมูลจากสื่อออนไลน์นั้น มีข้อมูลที่ไม่จำเป็นแฝงอยู่ จึงต้องการทำความสะอาดข้อมูล เช่น ลบ HTML tag ตลอดไปจนถึงเครื่องหมายที่ไม่จำเป็น หลังจากนั้นตัดคำแต่ละข้อความ แล้วเอาข้อความที่ตัดทำเป็นคลังคำศัพท์ โดยแต่ละคำศัพท์จะมี

ส่วนที่เป็นป้ายกำกับ ID แทนตัวศัพท์ และสร้าง feature vector ที่สามารถอธิบายลักษณะของคำศัพท์ได้

C. โมเดล

สถาปัตยกรรมของ Convolutional Neural Network โมเดลได้แรงบันดาลใจจากโมเดลการจำแนกประเภทของประโยคของ Yoon Kim[13] ซึ่งได้รับอิทธิพลจาก โมเดลของ Ronan Collobert[14] โดยกำหนดให้ $x_i \in R^k$ เป็นเวกเตอร์ความสอดคล้องกันของคำที่ตำแหน่ง i ของประโยคในเมตริก k มิติ และกำหนดให้

$$x_{1:n} = x_1 \oplus x_2 \oplus x_3 \oplus \ldots \oplus x_n \quad (1)$$

เมื่อ $\oplus$ เป็นตัวดำเนินการการต่อกันของคำ

ตัวดำเนินการคอนโวลูชันมีการเรียกใช้ฟิลเตอร์ $w \in R^{hk}$ ซึ่งจะถูกนำไปใช้ในแต่ละกรอบคำ h คำเพื่อสร้างฟีเจอร์ใหม่ โดยฟีเจอร์ c_i ถูกสร้างจากกรอบของคำ $x_{i:i+h-1}$ โดย

$$c_i = f(w \bullet x_{i:i+h-1} + b) \quad (2)$$

โดยที่ $b \in R$ คือตัวเลขที่นำเสนอความลำเอียงและ f คือฟังก์ชันที่ไม่ใช่ฟังก์ชันเชิงเส้น โดยฟิลเตอร์ถูกประยุกต์เข้าทุก ๆ ความเป็นได้ของกรอบของคำเพื่อสร้างแผนผังฟีเจอร์

$$c = [c_1, c_2, \ldots, c_{n-h+1}] \quad (3)$$

จากนั้นประยุกต์ใช้การดำเนินการ max-over-time pooling Ronan Collobert[14] เพื่อหาค่าที่มากที่สุดเป็นค่าความสอดคล้องกันของฟีเจอร์สำหรับฟิลเตอร์ใด ๆ

จากโมเดลการจำแนกประเภทของ Yoon Kim[13] ซึ่งสามารถจำแนกประเภทของประโยคซึ่งใช้ Softmax Classifier เป็นแอคทิเวชันฟังก์ชันของเอาต์พุตเลเยอร์ซึ่งสารมารแบ่งแยกทางสถิติออกเป็นคลาสได้ แต่ในการงานวิจัยมีคลาสเพียงแค่ 2 คลาส Sigmoid แอคทิเวชันฟังก์ชันจึงถูกนำมาใช้

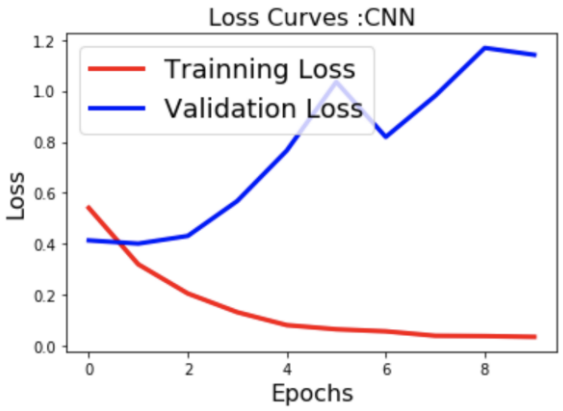
ตารางที่ I เลเยอร์และแอคทิเวชันฟังก์ชันของเครือข่ายการเรียนรู้ของเครื่อง

Input Layer	Input:	(None, 1000)
	Output:	(None, 1000)
Embedding	Input:	(None, 1000)
	Output:	(None, 1000, 100)
Conv1D Rectifier Activation Function	Input:	(None, 1000, 100)
	Output:	(None, 996, 128)
Max Pooling 1D	Input:	(None, 996, 128)
	Output:	(None, 199, 128)
Conv2D Rectifier Activation Function	Input:	(None, 199, 128)
	Output:	(None, 195, 128)
Max Pooling 2D	Input:	(None, 195, 128)
	Output:	(None, 39, 128)
Conv3D Rectifier Activation Function	Input:	(None, 39, 128)
	Output:	(None, 39, 128)
Max Pooling 3D	Input:	(None, 35, 128)
	Output:	(None, 1, 128)
Flatten	Input:	(None, 1, 128)
	Output:	(None, 128)
Dense Rectifier Activation Function	Input:	(None, 128)
	Output:	(None, 128)
Dense Sigmoid Activation Function	Input:	(None, 128)
	Output:	(None, 2)

IV. RESULTS

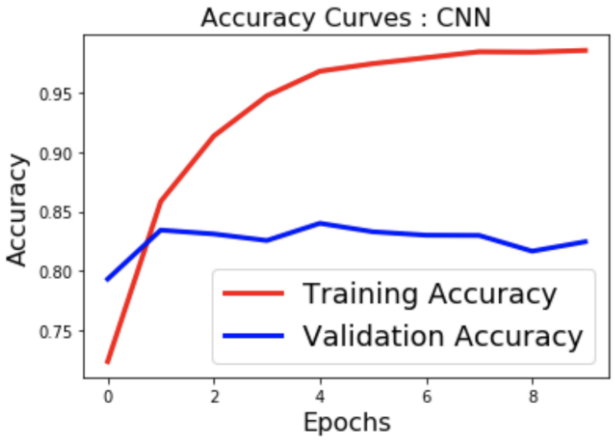
ชุดข้อมูลได้ถูกสร้างขึ้นจากความคิดเห็นภาษาอังกฤษในวีดีโอ 8154 ความคิดเห็นจาก 33 วีดีโอ แบ่งเป็นวีดีโอที่แสดงให้เห็นความรุนแรง 12 วีดีโอและไม่แสดงความรุนแรง 21 วีดีโอ อัตราส่วนความคิดเห็นวีดีโอความรุนแรง 36 %

จากการทดลองให้ผลลัพธ์ค่า Loss ดังนี้



รูปที่ 1 ค่า Loss Curves ของโมเดล

ค่าความแม่นยำการทำนายจากข้อมูลที่ใช้ในการเรียนรู้และข้อมูลที่ใช้ในการตรวจสอบมีค่าดังนี้



รูปที่ 2 ค่าความแม่นยำการทำนายของโมเดล

V. CONCLUSION

ผลลัพธ์จากการทดลองแสดงให้เห็นถึงความแม่นยำของการทำนายจากข้อมูลและโมเดลสามารถทำนายได้ว่าวีดีโอที่กำลังชมจะแสดงเนื้อหาที่มีความรุนแรงมีความแม่นยำสูงถึง 84% โมเดลไม่สามารถเพิ่มความสามารถด้วยการเรียนรู้ได้เมื่อเพิ่มรอบการเรียนรู้แต่ประสิทธิภาพอยู่ในขั้นพึงพอใจ ทั้งนี้ยังมีอักขระบางตัวที่ยังไม่สามารถแปลงเป็นเวกเตอร์ได้และอาจส่งต่อการทำนายที่ผิดพลาด

REFERENCES

[1] Phakawat Srakit, Thanuruk Theeramunkong, Choochart Haruechaiyasak, Manabu Okumura. “Classifying emotion in Thai YouTube comments”, 6th International Conference of Information and Communication Technology for Embedded System, 2015.

[2] Muhammad Zubair Asghar, Shakeel Ahmed, Afsana Marwat, Fazal Masud Kundi, “Sentiment Analysis on YouTube: A Brief Survey”, MAGNT Research Report, 2015.
[3] Abhidech Tepinm, Surangrut Jumnianpol, Ph.D, “Factors Affecting Levels of Violence from Social Action to the Different and Counterattack through Social Media”, Journal of Social Work, 2018.
[4] Noochanat Pinmuang, Jaree Thongkam, “Classifying Thai opinions on online media using text mining”, Journal of Science & Technology MSU, 2018.
[5] Niyanta Ashar, Hitarthi Bhatt, Shraddha Mehta, Chetashri Bhadane, “A Framework for Detection of Video Spam on YouTube”, International Journal of Computer Science and Information Technologies, 2015.
[6] Araya Giatgong, Kamolsak Wongsrikweak, Piyawat Giatgong, “Effects and Solutions of the Use of Online Social Network in Youths”, Journal of Criminology and Forensic Science, 2019.
[7] Túlio Alberto, Johannes Lochter, Tiago Almeida. “Tubespam: Comment Spam Filtering on YouTube”. In IEEE ICMLA, 2015.
[8] Rajeshwari Kandakatla, “Identifying Offensive Videos on YouTube”, B.Tech., Kakatiya University, 2014.
[9] https://developers.google.com/youtube, Retrieved October 14, 2019
[10] Lakshmish Kaushik, Abhijeet Sangwan, John H.L. Hansen, “Automatic Sentiment Extraction from YouTube Videos”, IEEE, 2013.
[11] https://www.nltk.org/api/nltk.sentiment.html, Retrieved October 14, 2019
[12] https://creatoracademy.youtube.com/page/lesson/policy-violence, Retrieved October 26, 2019
[13] Yoon Kim, “Convolution Neural Networks for Sentence Classification”, New York University, 2011.
[14] Ronan Collobert, Jason Weston, Leon Buttou, Michael Karlen, Koray Kavukcuoglu, Pavel Kuksa, NEC Laboratories, “Natural Language Processing (Almost) from Scratch”, Journal of Machine Learning Research 12, 2011.

Analysis of touristic places in Thailand from Facebook

Pakpoom Vichianroj, Suppanut Nuangnidnaraporn, Achara Charoentanaworakun
Department of Computer Engineering
Faculty of Engineering
Chulalongkorn University
Bangkok, Thailand
{6270210321, 6270282521, 6270315621}@student.chula.ac.th

***บทคัดย่อ*—หัวข้อที่ได้รับความนิยมจากผู้ใช้โซเชียลมีเดียในประเทศไทยคือการรีวิวสถานที่ท่องเที่ยวในประเทศไทย การวิจัยนี้ได้เห็นประโยชน์ของข้อมูลในการใช้โซเชียลมีเดียดังกล่าว จึงได้วิเคราะห์ความนิยมสถานที่ท่องเที่ยวจากเฟซบุ๊กเพจโดยใช้เทคนิคต่าง ๆ ในการเก็บรวบรวมและวิเคราะห์ข้อมูล เพื่อให้สามารถนำข้อมูลที่ได้จากการวิเคราะห์มาต่อยอดในเชิงการตลาดสำหรับการกระตุ้นเศรษฐกิจของพื้นที่ใกล้เคียงของสถานที่ท่องเที่ยวนั้น ๆ ได้**

***Abstract*—A popular topic on social media in Thailand is the review of touristic places. This research focuses on the use of data on social media. This work analyses the popularity of touristic places from Facebook. Various data collecting techniques are used and the data analysis is used to marketing in order to improve the economic situation of touristic places.**

Keywords— sentiment analysis, social analysis, Facebook

I. INTRODUCTION

ในอดีตการเข้าถึงข้อมูลด้านการท่องเที่ยวนั้นมีอยู่จำกัด โดยทั่วไปสามารถเข้าถึงข้อมูลได้เพียงผ่านทางสื่อสิ่งพิมพ์ต่าง ๆ อาทิเช่น หนังสือพิมพ์ หนังสือแนะนำการท่องเที่ยวตามสถานที่ต่าง ๆ เป็นต้น ข้อมูลที่ได้รับจึงอยู่ในวงแคบ และไม่สามารถเข้าถึงผู้คนได้ทุกระดับ รวมถึงความหลากหลาย และแง่มุมของการนำเสนอข้อมูลยังมีอยู่อย่างจำกัด แต่ในปัจจุบันด้วยความนิยมของโซเชียลมีเดีย (Soical Media) ต่าง ๆ เช่น เฟซบุ๊ก (Facebook) อินสตราแกรม (Instagram) ทวิตเตอร์ (Twitter) เป็นต้น ทำให้การเข้าถึงข้อมูลได้ง่ายขึ้น มีความน่าเชื่อถือมากขึ้นเรื่อย ๆ รวมถึงสามารถสืบค้นข้อมูลและศึกษาเพิ่มเติมได้โดยง่าย

เฟซบุ๊กเป็นโซเชียลมีเดียที่ได้รับความนิยมมากที่สุดทั่วโลก รวมถึงในประเทศไทย การใช้เฟซบุ๊กสำหรับสังคมไทยในปัจจุบันนอกเหนือไปจากการใช้เพื่อการติดต่อสื่อสาร เผยแพร่ และติดตามความเคลื่อนไหว รวมทั้งแสดงความคิดเห็นระหว่างกลุ่มเพื่อน และกลุ่มสังคมในเฟซบุ๊กแล้ว อีกอย่างหนึ่งที่ได้รับความนิยมไม่แพ้กัน คือ การใช้เฟซบุ๊กในการเขียน และอ่านรีวิว (Review) สิ่งต่าง ๆ ที่ได้รับความนิยม หรือกำลังเป็น กระแส ซึ่งตรงกับความสนใจของผู้ใช้เฟซบุ๊กแต่ละคน โดยหลาย ๆ คนนอกจากทำรีวิวเพื่อเป็นการแบ่งปันข้อมูลกับเพื่อน ๆ ในเฟซบุ๊กแล้ว ยังมีการเปิดเพจสาธารณะเพื่อให้ผู้คนทั่วไปที่สนใจเข้าถึงข้อมูลต่าง ๆ ได้เช่นเดียวกัน

สถานที่ท่องเที่ยวในประเทศไทยก็เป็นสิ่งหนึ่งที่ผู้ใช้เฟซบุ๊กนิยมนำมาเป็นหัวข้อในการทำรีวิวเป็นจำนวนมาก ทั้งนี้เพราะสถานที่ท่องเที่ยวในประเทศไทยมีอยู่ในทุกจังหวัด และสถานที่ท่องเที่ยวในแต่ละจังหวัดแต่ละสถานที่มีความสวยงาม เสน่ห์ และเอกลักษณ์ที่แตกต่างกันออกไป นอกเหนือจากแง่มุมความสวยงามของแต่ละสถานที่ท่องเที่ยวแล้ว ปัจจัยอื่นที่เกี่ยวข้องกับการท่องเที่ยว ก็เป็นหัวข้อที่เป็นส่วนหนึ่งในการทำรีวิว และผู้คนที่เข้ามาอ่านรีวิวก็ให้ความสนใจ เช่น วิธีการเดินทาง สถานที่พักในบริเวณใกล้เคียง ร้านอาหาร เป็นต้น

พฤติกรรมของผู้ใช้เฟซบุ๊กส่วนใหญ่จะมีการตอบสนอง และเข้าไปมีส่วนร่วมกับโพสต์ที่เกี่ยวข้องกับสถานที่ที่ผู้ใช้เฟซบุ๊กชื่นชอบ หรือเป็นสถานที่ที่ตนเองสนใจอยากที่จะไป จึงทำให้ผู้วิจัยสนใจศึกษา และวิเคราะห์ความนิยมสถานที่ท่องเที่ยวจากเพจสาธารณะ ซึ่งมีความน่าเชื่อถือ และมีผู้คนที่เข้าถึงเป็นจำนวนมาก เพื่อเป็นประโยชน์ในด้านการทำการตลาดของภาคส่วนรัฐบาลและเอกชน และเป็นประโยชน์ของผู้คนท้องถิ่นของสถานที่ท่องเที่ยวต่าง ๆ ที่จะสามารถใช้ข้อมูลดังกล่าว มาใช้เพื่อการปรับปรุงเปลี่ยนแปลง สินค้าและบริการของตนเอง ทำให้เกิดการกระตุ้นเศรษฐกิจของคนในพื้นที่ได้

การวิจัยในฉบับนี้ได้ทำการรวบรวมข้อมูลจากเพจเฟซบุ๊กสาธารณะที่เกี่ยวข้องกับการท่องเที่ยวที่ได้รับความนิยมในประเทศไทย โดยการเก็บข้อมูลของสถานที่ท่องเที่ยวต่าง ๆ ด้วยจำนวนผู้ที่กดถูกใจ จำนวนคอมเมนต์ (Comment) และเนื้อหาที่ปรากฏในคอมเมนต์ จากนั้นนำสถานที่ท่องเที่ยวแต่ละสถานที่นำมาจัดกลุ่มตามประเภทของสถานที่ แล้วนำข้อมูลผู้ที่กดถูกใจ และจำนวนคอมเมนต์ มาวัดความนิยมของสถานที่ท่องเที่ยวประเภทต่าง ๆ

II. LITERATURE REVIEW

สถานที่ท่องเที่ยวในประเทศไทยสามารถแบ่งออกได้หลายประเภทตามกิจกรรม ซึ่งได้กล่าวไว้ในงานวิจัย [7] โดยงานวิจัยดังกล่าวได้อ้างอิงข้อมูลจากการท่องเที่ยวแห่งประเทศไทย ยกตัวอย่างเช่น ผจญภัย เดินป่า ปีนเขา ท่องเที่ยวเชิงวัฒนธรรม เป็นต้น ในงานวิจัย [2] ได้ใช้ K-Nearest Neighbor Algorithm (K-NN) ในการจัดกลุ่มสถานที่ท่องเที่ยว ที่มีกลุ่มของข้อมูลใกล้เคียงกัน และใช้ Semantic Web Rule Language (SWRL) ในการสร้างกฎที่ใช้ในการแนะนำสถานที่ท่องเที่ยวจากกลุ่มประเภทของสถานที่ให้แต่ละบุคคล โดยใช้ตัวแปรต่าง ๆ คือ อายุ เพศ ความสนใจ และประเภทสถานที่ท่องเที่ยวตามกิจกรรม

มีผลการวิจัยที่ศึกษาเกี่ยวกับใช้โซเชียลมีเดียที่เกี่ยวข้องกับการท่องเที่ยวและสถานที่ท่องเที่ยวในประเทศไทย โดยเป็นการศึกษาว่ามีความสัมพันธ์ระหว่างการใช้ และความเชื่อถือในสื่อโซเชียลกับการท่องเที่ยวหรือไม่ [3] จากการศึกษาพบว่า มีการใช้สื่อโซเชียลในก่อน ระหว่าง และภายหลังการท่องเที่ยว เป็นไปในการเตรียมการท่องเที่ยว หาแรงบันดาลใจ และหาข้อมูลเกี่ยวกับจุดหมายปลายทาง และสถานที่พัก ซึ่งสื่อโซเชียลที่ถูกใช้มากที่สุด คือ สื่อโซเชียลของบุคคลที่สาม

ในการวัดความนิยมของผู้ใช้โซเชียลมีเดีย สามารถใช้รูปแบบการวัดหลากหลายแบบในการพิจารณา เช่น จำนวนผู้ใช้ที่เป็นเพื่อน กดติดตาม กด

ถูกใจเพจ (Page) จำนวนผู้ใช้ที่เข้ามาดูข้อมูลอย่างสม่ำเสมอ จำนวนการกดถูกใจเนื้อหาที่แบ่งปัน จำนวนคอมเมนต์ของผู้ใช้ เป็นต้น โดยตัววัดแต่ละตัวย่อมมีความสำคัญและน้ำหนักในการพิจารณาความนิยมที่แตกต่างกันออกไป ในงานวิจัย [1] พบว่าตัววัดที่มีความสำคัญในการวัดตามลำดับได้แก่ การมีส่วนร่วมของผู้ใช้ การแผ่กระจายของข้อมูล ความรู้สึกร่วม ในขณะที่งานวิจัย [6] ได้นำข้อมูลการรีวิวบนเว็บไซต์มาทำการวิเคราะห์โดยใช้ การทำเหมืองข้อมูล (Data mining) และการประมวลผลภาษาธรรมชาติ (Natural Language Processing) มาช่วยในการสร้างกระบวนการนับข้อความ และความคิดเห็นที่เกี่ยวข้องโดยตรงกับสิ่งที่ทำการรีวิว โดยจะถือว่าข้อความที่ถูกกล่าวถึงเป็นจำนวนมากเป็นข้อความที่เกี่ยวข้อง แล้วจึงทำการแบ่งแยกว่าในแต่ละครั้งที่มีการกล่าวถึงเป็นการกล่าวถึงในแง่บวก หรือลบ

III. METHODOLOGY

A. ต้นไม้ตัดสินใจ (Decision Tree)

ต้นไม้ตัดสินใจ (Decision Tree) [8] เป็นโมเดลที่ใช้ในการจำแนกประเภทข้อมูลของการเรียนรู้เครื่อง (Machine Learning) แบบหนึ่ง ซึ่งเป็นการจำแนกข้อมูลที่แบ่งตามคุณสมบัติข้อมูล ซึ่งโครงสร้างจะประกอบด้วย ราก (root) โหนด (node) กิ่ง (branch) และใบ (leaf)

งานวิจัยนี้ใช้ต้นไม้ตัดสินใจในการแบ่งข้อมูลประเภทของการท่องเที่ยวตามงานวิจัย [7] ซึ่งสามารถแบ่งประเภทการท่องเที่ยวออกเป็น 19 ประเภท ได้แก่ 1) ท่องเที่ยวผจญภัย 2) ปีนถ้ำ 3) ขี่นดอย 4) ขี่ช้าง 5) ปีนน้ำตก 6) ล่องแก่ง 7) ปั่นจักรยานเสือภูเขา 8) ดำน้ำดูปะการัง 9) พายเรือคายัก 10) โหนสลิง 11) กระโดดร่ม 12) ท่องเที่ยวเชิงการแพทย์ 13) ท่องเที่ยวเชิงธรรมชาติ 14) ท่องเที่ยวเชิงวัฒนธรรม 15) ท่องเที่ยวเชิงสุขภาพ 16) ท่องเที่ยวเชิงอบรม 17) ล่องเรือ 18) ท่องเที่ยวเชิงกีฬา 19) ท่องเที่ยวเชิงการศึกษา เพื่อนำข้อมูลมาในแต่ละประเภทมาเป็นตัวแปรในการวัดความนิยมในแต่ละประเภทของการท่องเที่ยว

B. การวิเคราะห์อารมณ์ (Sentiment Analysis)

การวิเคราะห์อารมณ์ (Sentiment Analysis) เป็นกระบวนการวิเคราะห์อารมณ์ และความรู้สึกของผู้ใช้ผ่านทางข้อความ โดยการวิเคราะห์อารมณ์เป็นการประมวลผลภาษาธรรมชาติ ซึ่งเป็นส่วนยอดนิยมในการเรียนรู้เครื่อง

งานวิจัยนี้จะวิเคราะห์อารมณ์ของผู้ใช้จากคอมเมนต์เฟซบุ๊กในโพสต์ของเฟซบุ๊กเพจ ว่ามีความรู้สึกต่อสถานที่ท่องเที่ยวนั้น ๆ อย่างไร เป็นความรู้สึกด้านบวกหรือด้านลบ และมีความสนใจต่อสถานที่นั้น ๆ มากน้อยเพียงใด แล้วนำมาแปลงค่าออกมาตัวเลข เพื่อใช้ในการวิเคราะห์ความนิยม

C. เพจแรงก์ (PageRank)

เพจแรงก์ (PageRank) [5] เป็นขั้นตอนวิธีจัดค่าตัวเลขบ่งบอกถึงความสำคัญของข้อมูลในกลุ่มของชุดข้อมูล โดยทางบริษัทกูเกิ้ล (Google) ใช้ในการจัดอันดับของเสิร์ชเอนจิน (search engine) ถ้ายิ่งคะแนนเพจแรงก์สูงก็จะแสดงไว้หน้าแรก ๆ ถ้าต่ำก็จะแสดงไว้หน้าในลำดับถัดไป โดยคำนวณจากสมการที่ (1)

$$PR(A) = (1-d) + d\,(PR(T_1)/C(T_1) + \ldots + PR(T_n)/C(T_n)) \quad (1)$$

งานวิจัยนี้ใช้เทคนิคเพจแรงก์ในการจัดอันดับความนิยมของการท่องเที่ยวแยกตามประเภทการท่องเที่ยว โดยนำจำนวนไลก์ จำนวนคอมเมนต์ และคะแนนการวิเคราะห์อารมณ์ มาใช้เป็นตัวแปร

IV. PROCEDURE

A. ชุดข้อมูล

- ประเภทการท่องเที่ยว ยกตัวอย่างเช่น เดินป่า ปีนเขา ท่องเที่ยวเชิงวัฒนธรรม เป็นต้น
- คะแนนการวิเคราะห์อารมณ์จากคอมเมนต์
- ข้อมูลจากเฟซบุ๊ก
  - ชื่อสถานที่
  - จำนวนคอมเมนต์ของแต่ละโพสต์
  - เนื้อหาคอมเมนต์ของแต่ละโพสต์
  - จำนวนไลก์ของแต่ละโพสต์

ตารางที่ I ตัวอย่างข้อมูล

ประเภทท่องเที่ยว	ชื่อสถานที่	จำนวนคอมเมนต์	คอมเมนต์	จำนวนไลก์	คะแนน
ท่องเที่ยวเชิงธรรมชาติ	ม่อนแจ่ม	361	{“เป็นสถานที่ที่สวยงามมาก ๆ ”, gender: ‘M’ age:’23’}]	15,000	0.69679 1247

B. ขั้นตอนวิธี

1. เก็บข้อมูลจากเพจเฟซบุ๊กที่เปิดสาธารณะเท่านั้นด้วยเครื่องมือจากการอ่านหน้าเว็บ และใช้เทคนิคการทำเหมืองข้อมูลเพื่อทำความสะอาดข้อมูล (Data cleaning) เพื่อลดความผิดพลาดของข้อมูล
2. นำรายละเอียดข้อมูลสถานที่ท่องเที่ยวมาแบ่งตามประเภทของสถานที่ท่องเที่ยว ด้วยขั้นตอนวิธีของต้นไม้ตัดสินใจ
3. รวบรวมคอมเมนต์จากแต่ละโพสต์มาคัดแยก และคัดเลือกเฉพาะคอนเมนต์ที่แสดงอารมณ์ต่อสถานที่นั้น ๆ เท่านั้นด้วยวิธีการประมวลผลธรรมชาติ โดยจะไม่นำข้อความที่เป็นการตั้งคำถาม หรือแท็กชื่อผู้คนมาใช้
4. นำคอมเมนต์จากแต่ละโพสต์หลังจากคัดแยกประเภทของคอมเมนต์ มาวิเคราะห์อารมณ์ต่อ สถานที่นั้น ๆ ว่ามีความรู้สึกเชิงบวกหรือเชิงลบ แล้วแสดงผลออกมาเป็นค่าตัวเลขเพื่อนำข้อมูลมาวิเคราะห์ความนิยมของสถานที่ท่องเที่ยว
5. นำจำนวนคอมเมนต์ จำนวนไลก์ คะแนนการวิเคราะห์อารมณ์ และประเภทการท่องเที่ยวมาจัดความนิยมด้วยเทคนิคเพจแรงก์ ตามประเภทการท่องเที่ยว

C. ผลการทดลอง

ตารางที่ II ผลการทดลอง

ประเภทการท่องเที่ยว	คะแนน	ลำดับ
ท่องเที่ยวเชิงธรรมชาติ	0.96249554	1
ล่องเรือ	0.88785646	2
ขี่ช้าง	0.77709928	3
ขึ้นดอย	0.68731178	4
ท่องเที่ยวผจญภัย	0.63056012	5
โหนสลิง	0.59312412	6
ท่องเที่ยวเชิงอบรม	0.5294589	7
ดำน้ำดูปะการัง	0.46914937	8
พายเรือคายัก	0.36898011	9
ท่องเที่ยวเชิงการแพทย์	0.30487469	10
ท่องเที่ยวเชิงสุขภาพ	0.29893621	11
กระโดดร่ม	0.24644307	12
ปั่นจักรยานเสือภูเขา	0.23876106	13
ท่องเที่ยวเชิงกีฬา	0.23179328	14
ปีนถ้ำ	0.21440466	15
ท่องเที่ยวเชิงการศึกษา	0.18285863	16
ปีนน้ำตก	0.15739424	17
ล่องแก่ง	0.08128913	18
ท่องเที่ยวเชิงวัฒนธรรม	0.05591258	19

V. CONCLUSION

สามารถจัดอันดับความนิยมของสถานที่ท่องเที่ยวจากเฟซบุ๊กเพจโดยใช้ต้นไม้ตัดสินใจ การวิเคราะห์อารมณ์ และอัลกอริทึมเพจแรงก์มาช่วยในการจัดกลุ่มข้อมูล วิเคราะห์คอมเมนต์ และเรียงลำดับความนิยมตามลำดับ

จากผลการทดลองประเภทของสถานที่ท่องเที่ยวที่ได้ความนิยมมากที่สุด คือ ท่องเที่ยวเชิงธรรมชาติ เนื่องจากในประเทศไทยมีสถานที่ท่องเที่ยวประเภทนี้อยู่เป็นจำนวนมาก มีอยู่ในทุกภาค ครอบคลุมทั่วทั้งประเทศ ทำให้สะดวกในการเดินทางและค้นหาที่พักในบริเวณใกล้เคียง ในแต่ละสถานที่ล้วนแล้วแต่มีความสวยงามและเอกลักษณ์ที่แตกต่างกันไป จึงน่าจะเป็นสาเหตุที่ทำให้ผู้ใช้เฟซบุ๊กมีความสนใจ และมีความนิยมต่อการท่องเที่ยวเชิงธรรมชาติมากที่สุด

REFERENCES

[1] Anuja Aroraa and Shivam Bansalb, 2019, Measuring social media influencer index- insights from facebook, Twitter and Instagram, Journal of Retailing and Consumer Services.

[2] Chakkrit.Snae Namahoot and Naruepo,tr Panawong, 2016, A Tourism Recommendation System for Thailand using Semantic Web Rule Language and K-NN algorithm, Information: An International Interdisciplinary Journal

[3] Chalida Techajirakul and Kriengsin Prasongsukarn, 2019, Applied social media and the effects of its use during vacation travel: A case study of Millennials in Thailand, Journal of Thai Interdisciplinary Research.

[4] Ghaidaa A. Al-Sultany and Asraa A. Abd Al-Ameer2, 2018, Locations Ranking using Page Rank Algorithm.

[5] Page, Lawrence; Brin, Sergey; Motwani, Rajeev; Winograd, Terry, 1998, The PageRank Citation Ranking: Bringing Order to the Web.

[6] Minqing Hu and Bing Liu, Mining and Summarizing Customer Reviews.

[7] Mohamed Ali Sharafuddin, 2015, Types of Tourism in Thailand, e-Review of Tourism Research (eRTR).

[8] Quinlan, J. R., 1986, Induction of decision tree

Sentiment Analysis of Twitter data to predict credibility of online shops

Apisara Saelim, Nithirun Numnonda,Thanasit Rithanasophon
Department of Computer Engineering
Faculty of Engineering
Chulalongkorn University
Bangkok, Thailand
{6270312721, 6270142321, 6270111921}@student.chula.ac.th

***บทคัดย่อ*—เพื่อหาความความน่าเชื่อถือของร้านค้าหรือบัญชีผู้ใช้ ผู้วิจัยจึงนำเทคนิค Sentiment Analysis มาใช้ในการวิเคราะห์คำหรือข้อความที่กล่าวถึงบัญชีต่าง ๆ และทำการ Classification ด้วยโมเดล Naïve Bayes และ Support Vector Machine ทำให้ได้มาซึ่งโมเดลที่สามารถนำมาทำนายความน่าเชื่อถือของร้านค้าออนไลน์ได้ โดยงานวิจัยนี้จะทำการวิเคราะห์เฉพาะร้านค้าภายในประเทศไทยเท่านั้น แต่จะใช้ทั้งคำภาษาไทยและภาษาอังกฤษในการวิเคราะห์ ซึ่งผลจากการทดลองพบว่าโมเดล Naïve Bayes มีความแม่นยำที่สูงกว่า Support Vector Machine ในการทำนายว่าร้านค้าออนไลน์ใดที่ไม่มีความน่าเชื่อถือ**

***Abstract*— To assess the credibility of online shops, we use sentiment analysis on comments that mention the shop. Two classification techniques: Naïve Bayes and Support Vector Machine are used to predict the credibility of the online shops. This work focuses only Thai shops. The result shows that Naive Bayes is more accurate than Support Vector Machine.**

Keywords— Twitter, online shopping, sentiment analysis, Naïve Bayes, Support Vector Machine

I. INTRODUCTION

ปัจจุบันที่เทคโนโลยีได้ถูกพัฒนาไปอย่างก้าวกระโดด ทำให้การเชื่อมต่อกันนั้นสามารถ ทำได้ง่ายขึ้น ด้วยการเข้ามาของเทคโนโลยีอินเทอร์เน็ตจึงเกิดขึ้นเป็นสังคมออนไลน์อย่างที่รู้จักกัน ในชื่อของโซเชียล มีเดีย (Social Media) ซึ่งทวิตเตอร์ (Twitter) ถือเป็นอีกหนึ่งสังคมออนไลน์ ที่ได้รับความนิยมเป็นอย่างมากโดยเฉพาะในด้านของการกระจายข้อมูลข่าวสารได้อย่างรวดเร็วและทันกาล

ทำให้การดำเนินงานธุรกิจจะต้องมีการปรับตัวและเปลี่ยนแปลงตามไปด้วย โดยการใช้อินเตอร์เน็ตหรือโซเชียล มีเดีย (Social Media) เป็นเครื่องมือทางการตลาด เพื่อเข้าถึงกลุ่มเป้าหมายได้อย่างรวดเร็ว จึงเกิดช่องทางการซื้อขายแบบใหม่ในสังคมออนไลน์ ที่เรียกว่า ร้านค้าออนไลน์ (Online Shopping)

ผู้วิจัยจึงสนใจที่จะนำข้อมูลที่เกิดขึ้นบนทวิตเตอร์มาวิเคราะห์หาความน่าเชื่อถือของร้านค้าออนไลน์ ว่าน่าเชื่อถือมากน้อยเพียงใด โดยการนำข้อมูลจากการโพส รีทวีต หรือการถูกกล่าวถึง (Mention) มาวิเคราะห์เพื่อหาความความน่าเชื่อถือของบัญชีผู้ใช้ (account) ผู้ซึ่งเป็นเจ้าของร้านค้าออนไลน์บนทวิตเตอร์ โดยใช้เทคนิค Naïve Bayes [1] และ Support Vector Machine [1] ในการทำ Classification ของคำหรือข้อความ ว่าเป็นบวก (Positive) หรือ ลบ (Negative)

II. RELATED WORK

งานวิจัยที่เกี่ยวข้องกับการทำวิจัยนี้ สามารถแบ่งได้เป็น 2 ส่วนคือ ส่วนของการเก็บรวบรวมข้อมูล และ การวิเคราะห์ข้อมูล

A. การเก็บรวบรวมข้อมูล

งานวิจัยของ Pak and Paroubek [9] ได้มีการนำข้อมูลทวิตเตอร์มาสร้างเป็นคลังข้อมูลเพื่อนำมาวิเคราะห์ความคิดเห็นและความรู้สึก โดยใช้เอพีไอของทวิตเตอร์ (Twitter API) ในการเก็บรวบรวมข้อมูลเชิงบวกด้วยอิโมติคอน “ :) , :-) , =) , :D ” ข้อมูลเชิงลบด้วยอิโมติคอน “ :(, :-(, =(, ;(” รวมถึงข้อมูลที่ไม่แสดงความรู้สึกหรือเป็นกลาง ซึ่งค้นหาจากบัญชีของหนังสือพิมพ์ยอดนิยม เช่น นิวยอร์กไทมส์, วอชิงตันโพสต์ และอื่นๆ สามารถรวบรวมได้ 44 ฉบับ จำนวน 300,000 ทวีต

งานวิจัยของ E. Kouloumpis [10] ได้ทำการตรวจสอบความรู้สึกจากข้อความในทวิตเตอร์ โดยการประเมินจากคำศัพท์และคุณสมบัติต่างๆ และได้เก็บรวบรวมข้อมูลจากคลังข้อมูลที่แตกต่างกัน 3 แหล่ง ดังนี้ ชุดที่ 1 คือ ชุดข้อมูลแฮชแท็ก ซึ่งเป็นข้อมูลที่เกิดจากการวิเคราะห์การกระจายข้อมูลของแฮชแท็กจากคลังข้อมูลทวิตเตอร์เอดินบะระ (Edinburgh Twitter Corpus) ที่บรรจุทวีตจำนวน 97 ล้านทวีตในระยะเวลา 2 เดือน โดยจะกรองทวีตที่ซ้ำ ทวีตที่ไม่มีแฮชแท็ก และไม่ใช่ภาษาอังกฤษออก แล้วนำมาพิจารณาความถี่ของแฮชแท็กที่เกิดในเชิงบวก เชิงลบ และเป็นกลาง จากนั้นนำแฮชแท็กไปค้นหาทวีตเพื่อนำมาใช้ในการฝึกสอน (Training) ต่อไป ข้อมูลชุดที่ 2 คือ ชุดข้อมูลแสดงอารมณ์ ซึ่งถูกสร้างมาจาก Go, et al. [8] โดยเก็บรวบรวมข้อมูลเชิงบวกด้วยอิโมติคอน “:)” เชิงลบด้วยอิโมติคอน “:(“. และจะไม่พิจารณาจากข้อความที่ปรากฏทั้งเชิงบวกและเชิงลบ ชุดข้อมูลที่ 3 คือ ชุดข้อมูล iSieve ที่บรรจุทวีตที่ถูกรวบรวมและเขียนด้วยมือ (Hand-annotated) จากบริษัท iSieve ประมาณ 4,000 ทวีต โดยจะเลือกเพียงบางหัวข้อ และในแต่ละทวีตจะต้องสะท้อนให้เห็นถึงความรู้สึกเชิงบวก เชิงลบ และเป็นกลาง

B. การวิเคราะห์ข้อมูล

งานวิจัยของ C. Castillo. [11] ได้วิเคราะห์ความน่าเชื่อถือของข้อมูลหรือข่าวสารที่แพร่กระจายผ่านทางทวิตเตอร์ โดยแยกหัวข้อออกเป็น 2 ประเภท คือ ข่าวและการสนทนา โดยนำเสนอคุณลักษณะที่ใช้วิเคราะห์ความน่าเชื่อถือ ดังนี้ คุณลักษณะของข้อความ (Message-based features) คุณลักษณะของผู้ใช้ (User-based features) คุณลักษณะของหัวข้อ (Topic-based features) และคุณลักษณะของการเผยแพร่ข้อมูล (Propagation-based features) โดยงานวิจัยนี้ได้คัดเลือกคุณลักษณะที่ดีที่สุด (Best-feature selection) มา 15 คุณลักษณะจากคุณลักษณะทั้ง 4 ซึ่งจากการวิเคราะห์ด้วยคุณลักษณะดังกล่าว พบว่า ข่าวที่น่าเชื่อถือจะถูก

เผยแพร่ผ่านทางผู้เขียนที่มีจำนวนข้อความและจำนวนรีทวีตเป็นจำนวนมาก

งานวิจัยของ E. Kouloumpis. [10] ได้ทำการสำรวจวิธีการในการสร้างชุดข้อมูลฝึกสอน (Training data) เพื่อที่จะสามารถนำมาระบุหัวข้อได้รวดเร็วมากขึ้น โดยใช้แฮชแท็กในการระบุว่าเป็นข้อมูลเชิงบวก เชิงลบ หรือเป็นกลาง เพื่อใช้สำหรับการฝึกสอนข้อมูล โดยเริ่มจากการทดสอบการแบ่งกลุ่มจากชุดข้อมูลแฮชแท็ก (HASH) ชุดข้อมูลเกี่ยวกับอารมณ์ (EMOT) และชุดข้อมูล iSieve ด้วยวิธีคุณลักษณะเอ็นแกรม (n-gram features) คุณลักษณะเล็กซิคอน (Lexicon features) คุณลักษณะชนิดของคำ (Part-of-speech features) และคุณลักษณะไมโครบล็อก (Micro-blogging features) ทั้งนี้เมื่อพิจารณาการประเมินประสิทธิภาพของชุดข้อมูลโดยวัดจากค่าเฉลี่ยของค่าวัดประสิทธิภาพ (F-measure) และค่าความถูกต้อง (Accuracy) พบว่า วิธีที่ดีที่สุด คือ การใช้คุณลักษณะเอ็นแกรม, คุณลักษณะเล็กซิคอน และคุณลักษณะไมโครบล็อก ร่วมกับชุดข้อมูลแฮชแท็ก แต่เมื่อรวมวิธีคุณลักษณะชนิดของคำ พบว่าทำให้ประสิทธิภาพลดลง จึงสามารถพิสูจน์ได้ว่าการใช้แฮชแท็กในการฝึกสอนข้อมูลเป็นประโยชน์พอกับการรวบรวมข้อมูลเชิงบวก เชิงลบ แต่ทั้งนี้ข้อมูลที่ใช้ฝึกสอนอาจขึ้นอยู่กับประเภทของ การวิเคราะห์ข้อมูลด้วย

III. METHODOLOGY

งานวิจัยนี้ ได้ทำการเก็บข้อมูลโดยการใช้เครื่องมือที่เรียกว่า การค้นหาข้อมูลทวิตเตอร์ขั้นสูง (Twitter Advanced Search) จากเว็บไซต์ https://twitter.com/search-advanced ซึ่งจะเป็นการค้นหาข้อมูลและแสดงผลออกมาในหน้าเว็บเบราว์เซอร์ หลังจากนั้นจึงทำการเขียนโปรแกรมภาษา python โดยใช้ library selenium และ beautifulsoup4 ในการดึงข้อมูลออกมาจากหน้าเว็บเบราว์เซอร์ โดยจะทำการเก็บข้อมูลหลายประเภท เช่น กลุ่มคำ บุคคล สถานที่ หรือช่วงวันเวลาที่ข้อมูลถูกทวีต โดยจะค้นหาจากกลุ่มคำที่เกี่ยวกับการซื้อขาย เช่น Sale Shop Promo Deal Pre-order โกง เป็นต้น

โดยการเก็บรวบรวมข้อมูลนั้นจะถูกแบ่งเป็น 2 ชุดข้อมูล คือ ชุดข้อมูล tweet message ของร้านค้าทั่วไป และ ชุดข้อมูล tweet message ของร้านค้าที่มีการโกง ซึ่งลักษณะของตารางข้อมูลที่เก็บในแต่ละชุดข้อมูล จะถูกเก็บและแบ่งออกเป็น 2 ตาราง ได้แก่ ตารางข้อมูล tweet message และ ตารางข้อมูลทั่วไปของแต่ละบัญชี

โดยในส่วนของ รายละเอียดตารางข้อมูล tweet message ของทั้งร้านค้าทั่วไปและร้านค้าที่มีการโกง จะมีโครงสร้างข้อมูล คือ วันที่และเวลาที่ทวีต, ข้อความที่ทวีต, แฮชแท็ก, จำนวน Retweet ข้อความ, จำนวน Favorite ข้อความ และ ชื่อบัญชีที่ทวีตข้อความ ซึ่งข้อมูลที่กล่าวถึงร้านค้าทั่วไป มีจำนวน 6,266 records จากทั้งหมด 210 ร้านค้า ดังตัวอย่างข้อมูล แสดงใน Fig. 1

datetime	10/26/2019 1:54:11 PM
message	#พร้อมส่ง Loccitane immortelle reset 30ml ของแท้ห้วเองจากคิงพาวเวอร์ 1,500ส่งฟรีems
mention	@jnineone
retweet	27
favouriteg	13
tweet_by_user_id	@Prim_____

Fig. 1. ตัวอย่างตารางข้อมูล tweet message ของร้านค้าทั่วไป

ในขณะที่ข้อมูลที่มีคนกล่าวถึงร้านค้าที่มีการโกง มีจำนวน 2,320 records จากทั้งหมด 60 ร้านค้า ดังตัวอย่างข้อมูลแสดงใน Fig. 2

datetime	10/18/2019 7:22:03 AM
message	เตือนภัยค่ะ อินี่โกง ชื่อบช @yu_108shop เซรั่มจากณกลิ่นนน้ำหอม วัฒนา สำเภาทอง ยุวลักษณ์ น้ำดอกไม้ #กล้องฟิล์มราคาถูก
mention	@yu_108shop
retweet	10
favouriteg	1
tweet_by_user_id	@luta2026

Fig. 2. ตัวอย่างตารางข้อมูล tweet message ของร้านค้าที่มีการโกง

ส่วนของ รายละเอียดตารางข้อมูลทั่วไปของแต่ละบัญชี จะเก็บโครงสร้างข้อมูล คือ ชื่อบัญชี (ScreenName), ชื่อผู้ใช้ (user_id), จำนวนผู้ที่กำลังติดตามเจ้าของบัญชี (Follower) และ จำนวนผู้ที่เจ้าของบัญชีกำลังติดตาม (Following)

โดยวิธีการในการเก็บข้อมูลนั้น จะใช้ hashtag #โกง เพื่อเก็บข้อมูลรายชื่อร้านค้าที่มีการโกง และใช้ hashtag #พร้อมส่ง, #ลดราคา และ hashtag ต่าง ๆ เกี่ยวกับการซื้อขายสินค้าออนไลน์ ในการเก็บข้อมูลรายชื่อร้านค้าทั่วไป เมื่อเก็บรวบรวมข้อมูลทั้งหมดแล้วจึงนำไปเข้าสู่กระบวนการวิเคราะห์ แต่เนื่องจากข้อมูลที่ได้ทำการเก็บมาส่วนใหญ่เป็นภาษาไทย จึงจำเป็นต้องใช้ library pythainlp ในการจัดการกับข้อความ ก่อนที่จะนำไปประมวลผล ซึ่งวิธีการที่ใช้คือ การทำ Sentimental Analysis กับข้อมูลที่คนกล่าวถึงร้านค้าว่า มีค่าความคิดเห็นไปทางลบหรือทางบวก แล้วนำไป Classification เพื่อนำไปทำนายผลลัพธ์ร้านค้านั้นว่ามีเป็นร้านค้าทั่วไปหรือร้านค้าที่โกง

IV. DATA ANALYSIS

ส่วนของการวิเคราะห์ข้อมูลนั้นสามารถแบ่งออกได้เป็น 4 ส่วน โดยเริ่มจากการทำ Descriptive analysis คือ การนำข้อความหรือประโยคที่ได้จากการเก็บรวบรวมมาหาความถี่ของคำ ทั้งจากร้านที่มีความน่าเชื่อถือและร้านที่ไม่มีความน่าเชื่อถือ เพื่อเป็นการเตรียมข้อมูลไปทำการวิเคราะห์ต่อในส่วนของการทำ Sentiment Analysis หลังจากนั้น เมื่อได้ผลแล้วว่ากลุ่มคำใดเป็นบวกหรือลบ จึงนำข้อมูลไปทำการวิเคราะห์ต่อเพื่อหาความสัมพันธ์ระหว่างร้านที่น่าเชื่อถือและร้านที่โกง และในลำดับสุดท้ายจะเป็นเปรียบเทียบโมเดล โดยการนำข้อมูลที่ได้มาจากวิธีการข้างต้นมาทำการ Classification ด้วยโมเดล Naïve Bayes และ Support Vector Machine เพื่อทำนายความน่าเชื่อถือของร้านค้า

A. Descriptive analysis

เมื่อได้ข้อมูลของคำที่แบ่งออกจากประโยคมาแล้ว เราจึงมาหาความถี่ของคำ (Word Frequency) โดยแบ่งเป็นสองส่วนได้แก่กลุ่มคำที่คนกล่าวถึงร้านค้าที่มีความน่าเชื่อถือ ดังตัวอย่างข้อมูล แสดงใน Fig. 3 และกลุ่มคำที่คนกล่าวถึงร้านค้าที่มีการโกง ดังตัวอย่างข้อมูล แสดงใน Fig. 4 ซึ่งจะพบว่าร้านค้าที่ไม่มีความน่าเชื่อถือนั้นมีกลุ่มคำที่คล้ายกัน เช่นเดียวกับร้านค้าที่มีความน่าเชื่อถือ จึงนำข้อมูลเหล่านี้ไปทำเป็นพจนานุกรมสำหรับการวิเคราะห์ Sentiment Analysis

คำ	จำนวนคำ
ลดราคา	372
พร้อมส่ง	262
บาท	126
ถูกและดี	100
howtoperfect	98

Fig. 3. คำหรือกลุ่มคำต่าง ๆ ที่ได้จากการเก็บข้อมูลจากร้านค้าที่มีความน่าเชื่อถือที่มีจำนวนความถี่ของคำมากที่สุด 5 อันดับ

คำ	จำนวนคำ
โกง	468
โกงเงิน	92
คนโกง	69
เตือนภัย	64
ทวีตดีคนรีน้อย	29

Fig. 4. คำหรือกลุ่มคำต่าง ๆ ที่ได้จากการเก็บข้อมูลจากร้านค้าที่มีการโกงที่มีจำนวนความถี่ของคำมากที่สุด 5 อันดับ

B. Sentiment Analysis

การแยกประเภทของ Text หรือกลุ่มคำ สามารถทำได้หลายวิธี ซึ่งในงานวิจัยนี้จะใช้เทคนิค Support Vector Machine (SVM) [1] ซึ่งเป็นการทำ Classification ในรูปแบบ Supervised Learning โดยจะนำข้อมูลที่เก็บรวบรวมได้ จากขั้นตอนการทำ Data Collection และผ่านการทำการ Data Preprocessing ด้วยแล้วนั้น มา Cross Validation ทำการสอน (Train) สลับกับการทดสอบ (Test) ให้กับโมเดล (Model) เพื่อหาว่าข้อความ หรือ กลุ่มคำนั้น เป็นในเชิงบวก เชิงลบ

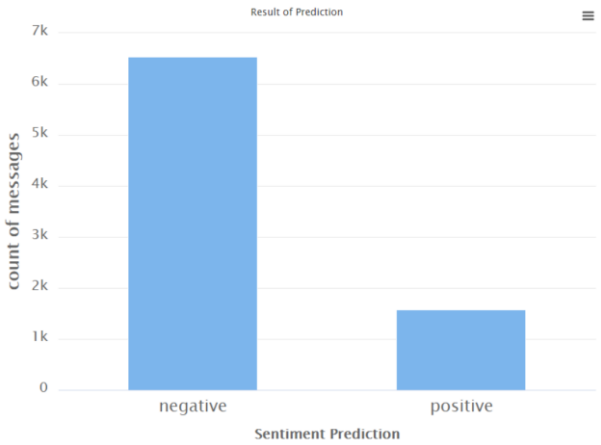


Fig. 5. กราฟแสดงผลการทำนายความรู้สึกกับข้อมูลทดสอบ

C. การวิเคราะห์ข้อมูล

เมื่อได้ข้อมูลจากการทำ Sentiment Analysis แล้วว่า ข้อความ หรือกลุ่มคำ ที่ผู้ใช้ทวีตนั้นเป็นไปในเชิงบวก เชิงลบ หรือ เป็นกลาง จึงจะนำข้อมูลเหล่านั้นมาทำการวิเคราะห์ Classification โดยการใช้ คุณลักษณะของข้อความ (Message-based features) คุณลักษณะของผู้ใช้ (User-based features) ประกอบกับค่าน้ำหนัก (Weight) ที่ได้จากผลลัพธ์การทำนาย Sentiment Analysis เพื่อหาความสัมพันธ์ระหว่าง ความน่าเชื่อถือของร้านค้ากับการถูกพูดถึง (Mention) ด้วยข้อความเชิงลบและข้อความเชิงบวก

D. การเปรียบเทียบโมเดล

จากการทดลองเอาข้อมูลมา Classification มาเพื่อแยกร้านค้าที่ไม่มีความน่าเชื่อถือกับร้านที่มีความน่าเชื่อถือ ด้วย Model Naïve Bayes และ Support Vector Machine พบว่า Naïve Bayes มีค่า class precision ในการทำนายร้านค้าที่มีความน่าเชื่อถือได้ 96.05% และมีค่า class precision ในการทำนายร้านค้าที่ไม่มีความน่าเชื่อถือได้ 65.49% ในขณะที่ Support Vector Machine มีค่า class precision ในการทำนายร้านค้าที่มีความน่าเชื่อถือได้ 81.99% และมีค่า class precision ในการทำนายร้านค้าที่ไม่มีความน่าเชื่อถือได้ 96.90%

แต่ในขณะเดียวกัน Naïve Bayes มีค่า class recall ในการทำนายร้านค้าที่มีความน่าเชื่อถือได้ 84.09% และ มีค่า class recall ในการทำนายร้านค้าที่ไม่มีความน่าเชื่อถือได้ 89.73% แต่ส่วน Support Vector Machine มีค่า class recall ในการทำนายร้านค้าที่มีความน่าเชื่อถือได้ 99.62% , มีค่า class recall ในการทำนายร้านค้าที่ไม่มีความน่าเชื่อถือได้ 34.94%

จากการทดลองพบว่าโมเดล Naïve Bayes นั้นมีค่า class prediction ในการทำนายร้านค้าที่ไม่น่าเชื่อถือต่ำกว่า โมเดล Support Vector Machine นั้นหมายความว่า โมเดล Naïve Bayes ที่ใช้ในการทำนายร้านค้าที่ไม่น่าเชื่อถือมีความถูกต้องน้อยกว่าโมเดล Support Vector Machine แต่ในอีกทางหนึ่ง โมเดล Naïve Bayes นั้นมีค่า class recall ในการทำนายร้านค้าที่ไม่น่าเชื่อถือสูงกว่าโมเดล Support Vector Machine นั้นหมายความว่า โมเดล Naïve Bayes มีการทำนายร้านค้าที่ไม่น่าเชื่อถือเป็นจำนวนที่มากกว่า Support Vector Machine

โดยรวมแล้ว Support Vector Machine มีค่าความแม่นยำอยู่ที่ 83.33% ในขณะที่ Naïve Bayes มีค่าความแม่นยำอยู่ที่ 85.51%

	true fair.csv	true cheat.csv	class precision
pred. fair.csv	2226.313	91.447	96.05%
pred. cheat.csv	421.13	799.274	65.49%
class recall	84.09%	89.73%	

Fig. 6. ผลการทดสอบความแม่นยำของ Model Naïve Bayes

	true fair.csv	true cheat.csv	class precision
pred. fair.csv	2637.481	579.462	81.99%
pred. cheat.csv	9.962	311.259	96.90%
class recall	99.62%	34.94%	

Fig. 7. ผลการทดสอบความแม่นยำของ Model Support Vector Machine

E. ผลการวิจัย

จากการทดลอง เราได้ผลลัพธ์เป็นโมเดล) Model (ที่จะใช้ในการทำนายข้อความใด ๆ ว่าข้อความที่กล่าวถึงนั้นมีคำในลักษณะเชิงบวกหรือเชิงลบ และข้อความที่เอ่ยถึงร้านค้านั้นเป็นร้านค้าที่มีความน่าเชื่อถือหรือไม่น่าเชื่อถือ ดังตัวอย่างข้อมูล แสดงใน Fig. 8

โดยโมเดลที่แนะนำให้นำมาใช้ คือ Naïve Bayes เนื่องจากมีความแม่นยำในการทำนายได้ถูกต้องมากกว่า ถึงแม้ว่าจะมีค่า class precision ที่ต่ำกว่า แต่เมื่อมาเฉลี่ยแล้วค่าความแม่นยำโดยรวมดีกว่า

Class	cheat.csv
prediction(Sentiment)	negative
confidence(negative)	0.724
confidence(positive)	0.276
message	เตือนภัย อิ โกง ชื่อ บช yu 108 shop เชษฐา ภณ กลิ่น น้ำหอ...

Fig. 8. ผลการทดสอบความแม่นยำของ Model Support Vector Machine

V. CONCLUSION

งานวิจัยจำนวนมากที่นำข้อมูลทวิตเตอร์มาวิเคราะห์ในมุมต่างๆ เช่น การวิเคราะห์ความน่าเชื่อถือของข้อมูลและข่าวสารที่แพร่กระจายในทวิตเตอร์ [11] หรือการนำข้อมูลทวิตเตอร์มาสร้างเป็นคลังข้อมูล เพื่อนำมาวิเคราะห์ความเห็นและความรู้สึก [9] ผู้วิจัยจึงได้เก็บรวบรวมข้อมูลทวิตเตอร์ที่เกี่ยวกับการซื้อขายสินค้าเพื่อนำมาวิเคราะห์ความน่าเชื่อถือของร้านค้าออนไลน์ โดยการสร้างโมเดลจากเทคนิค Support Vector Machine เพื่อใช้ในการทำนายความน่าเชื่อถือ

VI. FUTURE WORK

ผลลัพธ์ที่ได้มานั้นจะมีความแม่นยำของโมเดลที่สูงในระดับหนึ่ง แต่การตรวจจับหาร้านค้าที่มีความน่าเชื่อถือควรจำเป็นจะต้องมีความแม่นยำมากกว่านี้ เพื่อไม่ให้เกิดข้อผิดพลาดในการกล่าวหาร้านค้าที่มีความน่าเชื่อถือ ผู้วิจัยจึงแนวคิดที่จะพัฒนาโมเดลต่อไป โดยการเพิ่มคำ เพิ่มจำนวนของข้อมูล เพื่อพัฒนาโมเดลให้มีความแม่นยำมากยิ่งขึ้น

REFERENCES

[1] Abdullah Alsaeedi, and Mohammad Zubair Khan, “A Study on Sentiment Analysis Techniques of Twitter Data”, presented at (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 10 No. 2, 2019.

[2] Rushabh Shroff, and Amitash Ramesh, “Slang Word Identification on Twitter”, Amitash Rames et al, International Journal of Computer Technology & Applications,Vol7 No. 3, 2014.

[3] Xu Yongzhi, and Sun Lu, “Research on Consumer Confidence under the Online Shopping”, 2010 International Conference on Innovative Computing and Communication and 2010 Asia-Pacific Conference on Information Technology and Ocean Engineering, 2010.

[4] Yili Wang, KyungTae Kim, ByungJun Lee, and Hee Yong Youn, “Word clustering based On POS feature for efcient twitter sentiment analysis”, Wang et al. Hum. Cent. Comput. Inf. Sci. ,2018.

[5] Suresh Y, “Software quality assessment for open source software using logistic and Naïve Bayes classifer”, Paper presented at the International conference on computation system and information technology for sustainable solutions, Oct 2016.

[6] Mandy Johnson , Christie Bledsoe , Jodi Pilgrim, and Hollis Lowery-Moore, “Twitter: A Tool for Communities of Practice”, SRATE Journal, v28 n1 p Win 2019, 2019.

[7] Stephen Hil , and Rebecca Scott “Developing an Approach to Harvesting, Cleaning, and Analyzing Data from Twitter Using R”, Information Systems Education Journal (ISEDJ) 15 (3) ISSN: 1545-679X , May 2017.

[8] A. Go, R. Bhayani, and L. Huang., "Twitter sentiment classification using distant supervision," Technical report, Stanford Digital Library Technologies Project., 2009.

Prediction of the success and the number of goal in football game

Waruch Kaolikit, Krittayot Bherngjitt, Pichaiyut Siriprompisan
Department of Computer Engineering
Faculty of Engineering
Chulalongkorn University
Bangkok, Thailand
{6270248221, 6270008021, 6270186021}@student.chula.ac.th

***บทคัดย่อ*—การพยากรณ์ความสำเร็จในการยิงประตู และจำนวนประตูโดยการวิเคราะห์ข้อมูลแบบ Linear Regression และ Decision Tree โดยวัดความแม่นยำของเทคนิคต่าง ๆ เพื่อดูว่าวิธีการดังกล่าวเป็นวิธีการทำนายที่เหมาะสมหรือไม่โดยผลลัพธ์ได้คือวิธีการ Linear regression model สามารถใช้ทำนายจำนวนประตูที่นักเตะจะทำได้ และ แสดงให้เห็นตัวแปรที่มีผลในทางส่งเสริมกับจำนวนประตูที่ยิงได้คือ การใช้ศีรษะและตำแหน่งกองหน้า เป็นต้น Decision tree สามารถทำนายว่านักเตะจะยิงประตูเข้าหรือไม่ภายในสถานการณ์ต่าง ๆ ตัวแปรสถานการณ์ที่มีผลมากกับการยิงประตูได้ คือ การยิงตรงกรอบ การยิงในจังหวะ open play สามารถนำไปใช้ให้เกิดประโยชน์ในการฝึกฝนนักเตะ วางแผนการเล่น เพื่อสร้างให้เกิดโอกาสที่จะยิงประตูได้สูงขึ้น**

***Abstract*—The success and the number of goal in football game is predicted with two methods: linear regression and decision tree. The accuracy is used to find suitable methods. For predicting the number of goal, linear regression is suitable. The positive factors are the heading and the positive of forward. Decision tree however, can predict the success/failure of getting a goal. The variables are the accurate shoot to the frame, the shooting in open play. The prediction is used in the training, planning to improve the players.**

Keywords— football game prediction, linear regression, decision tree

I. บทนำ

ฟุตบอลเป็นกีฬาที่ผู้คนรู้จักให้ความสนใจติดตามเป็นจำนวนมากทั่วโลก ดังจะเห็นได้ตาม social network หรือรายการโทรทัศน์ ยิ่งไปกว่านั้นคือบรรยากาศของการรับชมที่มีกลุ่มของสังคมร่วมติดตามหรือร่วมเชียร์ทีมฟุตบอลที่ชื่นชอบไปด้วยกัน สิ่งต่าง ๆ เหล่านี้ทำให้กีฬาฟุตบอลเป็นที่สนใจของสังคมในวงกว้างประกอบกับในโลกยุคที่ข้อมูลข่าวสารสามารถเข้าถึงได้ง่ายบนพื้นฐานของต้นทุนการรับชมที่ต่ำลงยิ่งเป็นผลสนับสนุนให้ผู้คนสามารถเข้าถึงข่าวสารกีฬาและรับชมฟุตบอลได้ง่ายขึ้นกว่าก่อน อีกทั้งมันยังเป็นกีฬาอันดับหนึ่งที่เหล่านักเสี่ยงโชคลงเงินเพื่อการพนัน

ในการแข่งขันฟุตบอลผลแพ้ชนะจะถูกพิจารณาจากจำนวนประตูที่ยิงได้ กล่าวคือหากทีมใดสามารถยิงประตูได้มากกว่าก็จะเป็นฝ่ายชนะโดยในการที่จะได้ประตูมาแต่ละประตูนั้นผู้เล่นจะต้องอาศัยทักษะความสามารถทางด้านร่างกาย การฝึกฝน การทำงานเป็นทีม การคิดวิเคราะห์การแข่งขัน การวางแผนของทีม นอกจากการทำประตูจะเป็นตัวตัดสินผลแพ้ชนะของเกมแล้ว มันยังเป็นสิ่งที่ผู้ชมติดตามเกมและลุ้นไปกับการยิงประตูและการป้องกันไม่ให้อีกฝั่งยิงประตูได้ ดังนั้นการยิงประตูจึงเป็นสิ่งสำคัญที่สุดของการแข่งขันกีฬาฟุตบอลทั้งสำหรับผู้เล่นและผู้ชม ด้วยเหตุนี้กลุ่มผู้วิจัยจึงสนใจศึกษาหาปัจจัยที่ส่งผลต่อการยิงประตูและทำนายความสำเร็จในการยิงประตูในสถาณการณ์ต่าง ๆ จากสถิติของเหตุการณ์การทำประตูที่เกิดขึ้นในแต่ละการแข่งขัน อีกทั้งกลุ่มผู้วิจัยยังสนใจถึงการทำนายจำนวนประตูที่นักเตะคนหนึ่งจะยิงได้ต่อฤดูกาลจากสถิติการเล่นในอดีตของผู้เล่นคนนั้น ๆ เพื่อเป็นประโยชน์ต่อการพัฒนาเทคนิคการเล่นและเพิ่มอัตราการยิงประตูได้ของทีมฟุตบอลและยังสามารถใช้วิธีการทำนายดังกล่าวในการใช้เลือกนักเตะเข้าสู่ทีม รวมถึงเพิ่มความแม่นยำให้นักวิจารณ์ฟุตบอลซึ่งมักจะวิเคราะห์ให้ทัศนะและทำนายผลการแข่งขันให้มีความแม่นยำสูงขึ้น

ปัจจัยที่ส่งผลต่ออัตราหรือโอกาสที่จะยิงประตูได้นั้นสามารถแบ่งได้ออกได้เป็น 2 ประเภทหลักปัจจัยแรกคือปัจจัยเชิงเทคนิคในการเล่นกับลูกฟุตบอลและปัจจัยเชิงกายภาพ เช่น เท้าข้างที่ถนัด วิธีการยิงประตู ความแข็งแรงของร่างกาย ความแรงในการยิง ตำแหน่งในสนามที่ยิงประตู เป็นต้น ปัจจัยหลักที่สองคือสภาพแวดล้อมขณะที่ทำการแข่งขัน เช่น นาทีที่ทำประตูได้ จำนวนใบเหลืองที่ได้รับ จำนวนใบแดงที่ได้รับ เล่นเป็นทีมเหย้าหรือทีมเยือน เป็นต้น โดยศึกษาปัจจัยหลักทั้งสองด้านและประเมินความสำคัญของปัจจัยต่าง ๆ แล้วจึงสามารถพยากรณ์ความสำเร็จในการยิงประตูและจำนวนประตูที่นักเตะจะยิงได้ต่อฤดูกาลได้ผ่านเทคนิคทางการวิเคราะห์ข้อมูลรูป Linear Regression และ Decision Tree โดยวัดความแม่นยำของเทคนิคต่าง ๆ เพื่อดูว่าวิธีการดังกล่าวเป็นวิธีการทำนายที่เหมาะสมหรือไม่

II. ทบทวนวรรณกรรม

A. Choke or Shine - Quantifying Soccer Players Abilities to Perform Under Mental Pressure

Choke or Shine - Quantifying Soccer Players Abilities to Perform Under Mental Pressure ได้รับการตีพิมพ์ลงเว็บไซต์ เป็นส่วนหนึ่งของงาน “MIT SLOAN SPORTS ANALYTICS CONFERENCE RESEARCH PAPER”จาก MIT รอบ 8 คนสุดท้ายปี 2019 ซึ่งงานวิจัยนี้ว่าด้วยการวัดประสิทธิภาพในการเล่นของผู้เล่นแต่ละคนในสถานการตรึงเครียด เพื่อช่วยให้ทีมฟุตบอลตัดสินใจว่าจะซื้อตัวผู้เล่นคนไหนหรือผู้เล่นคนไหนมีจุดที่ต้องปรับปรุงในการเล่นในภาวะมีแรงกดดันสูง ซึ่ง paper นี้ได้ทำการประเมินสถานการตึงเครียดทั้งนอกเกม ยกตัวอย่างเช่น การแข่งขันครั้งนี้มีเป็นเกมที่มีความสำคัญมากเพราะชนะแล้วจะเข้ารอบหรือเป็นเกมที่เจอทีมคู่ปรับและในเกมเช่นการยิงประตูตอนใกล้หมดเวลา โดยในpaperนี้จะมีการวัดค่าประสิทธิภาพจากการตัดสินใจและการเล่นของผู้เล่นออกมาเป็นคะแนนโดยใช้ modelประเภท Gradient Boosted Ranking Trees โดยที่

ใช้ข้อมูลอ้างอิงจากสถิติของ Premiere League ของอังกฤษ, Laliga ของสเปน, และ Bundesliga ของเยอรมัน เป็นข้อมูลตั้งต้น และใช้ algorithm PyMC3 Auto-Differentiation Variational Inference ในการฝึก model โดยการเล่นในเกมจะแบ่งออกเป็น 3 ประเภทใหญ่ๆคือ ผลงานโดยรวมทั้งเกม, การตัดสินใจ, และการกระทำ เพราะผู้เล่นที่เล่นเก่งอาจจะตัดสินใจผิดพลาดและไม่ได้ทำการเล่นที่ดีที่สุดที่จะทำได้ในสถานการนั้นๆ โดยงานวิจัยนี้ทำให้เห็นรูปแบบการตัดสินใจของทีมเช่นทีมLiverpoolนั้นจะเล็งตัวนักเตะที่สามารถเล่นได้ดีในสภาวะตึงเครียด โดยที่คณะผู้วิจัยเห็นด้วยกับงานวิจัยนี้ว่าการวัดการเล่นของผู้เล่นในสถานการตึงเครียดนั้นมีความสำคัญ เพราะผู้เล่นที่เล่นได้ดีเมื่อมีแรงกดดันย่อมมีค่ามากกว่าผู้เล่นที่รับแรงกดดันไม่ไหว แต่งานวิจัยนี้จะเน้นวิเคราะห์การยิงประตูของผู้เล่นในทุกสถานการไม่ใช่เฉพาะที่ตึงเครียดแต่งานวิจัยนี้ทำให้คณะผู้วิจัยคิดว่าควรให้คะแนนกับการยิงประตูในสถานการที่ตึงเครียดมากขึ้น

B. Decomposing the Immeasurable Sport: A deep learning expected possession value framework for soccer

นอกจากนี้ยังมีงานวิจัยที่ค้นหาว่าการกระทำใดในระหว่างเกมส่งผลต่อเกมนั้นมากที่สุดเช่น Decomposing the Immeasurable Sport: A deep learning expected possession value framework for soccerที่ได้รับการตีพิมพ์ลงเว็บไซต์http://www.sloansportsconference.com และเป็นผู้เข้ารอบรอบ8คนสุดท้ายปี 2019 เช่นกัน มีการใช้ possession value (EPV) framework มาวิเคราะห์ว่าการตัดสินใจของผู้เล่นและแผนการของแต่ละทีมมีการส่งผลต่อเกมที่เล่นเพียงใดเช่น การส่งบอลหรือความเสี่ยงของแผนการในการเล่น

ผู้ที่ทำงานวิจัยนี้ได้สรุปออกมาว่าmodelที่ใช้ในงานวิจัยนี้สามารถนำไปปรับใช้กับการวิเคราะห์เชิงลึกทั้งในด้านประสิทธิภาพการเล่นของผู้เล่นแต่ละคนหรือแบบทีมและการกำหนดค่าให้กับการกระทำแต่ละอย่างของผู้เล่นเพื่อประเมินการตัดสินใจของผู้เล่น โดยที่คณะผู้วิจัยคิดว่างานวิจัยนี้จะสามารถเป็นแหล่งอ้างอิงในการกำหนดคะแนนของการยิงประตูในงานวิจัยนี้ได้โดยที่จะคิดเฉพาะการยิงประตูไม่ใช่ภาพรวมของเกม

C. Quality vs Quantity": Improved Shot Prediction in Soccer using Strategic Features from Spatiotemporal Data

งานวิจัยชิ้นนี้เป็นการวิเคราะห์การยิงประตูโดยใช้ปัจจัยทาง spatiotemporal patternหรือปัจจัยเชิงพื้นที่และเวลาในการประมาณความเป็นไปได้ในการทำประตูนั้นสามารถใช้ข้อมูลบริบทของสถานการณ์ในช่วงก่อนทำประตูเพื่อประมาณโอกาสในการได้ประตูให้ดีขึ้นได้โดยบริบทต่าง ๆ เหล่านั้น ได้แก่ ตำแหน่งของผู้เล่นตำแหน่งป้องกันของฝ่ายตรงข้ามแบบแผนและโครงสร้างที่ทีมตรงข้ามใช้ในการป้องกัน รูปแบบการโจมตีของฝ่ายบุก ค่าความคาดหวังในการได้ประตู (Expected Goal Value EGV) ซึ่งได้ผลออกมาว่านอกจากช่วงเวลาของเกมเช่นการเล่นปกติหรือการยิงลูกโทษยังมีปัจจัยอื่นๆเช่นตำแหน่งของผู้เล่นอื่นและตำแหน่งของจุดที่ยิงก็ส่งผลต่อการยิงประตูด้วย คณะผู้วิจัยคิดว่าควรนำปัจจัยที่มีผลต่อการยิงประตูจากงานวิจัยนี้มารวมเป็นปัจจัยวิเคราะห์การยิงประตูในงานวิจัยของคณะผู้วิจัยด้วย

III. แหล่งข้อมูล

การทำการวิจัยนี้เก็บ dataset ทั้งหมด 3 ชุด โดยทั้งหมดได้จาก www.kaggle.com ซึ่งเป็น web-site ที่เป็นแหล่งรวมข้อมูลจากหลากหลายที่ หลากหลายสาขา ให้ข้อมูลที่เปิดเผยต่อสาธารณะและไม่มีค่าใช้จ่ายแก่บุคคลทั่วไปให้สามารถนำไปใช้ประโยชน์วิเคราะห์ได้อย่างอิสระ

ข้อมูลชุดที่ 1 Statbunker Football Statistics : รวบรวมสถิติในสองด้าน คือ

1) วิธีและลักษณะการยิงประตูของนักฟุตบอลแต่ละคน

2) สถิติจำนวนใบเหลืองและใบแดงที่ได้รับของนักฟุตบอลแต่ละคน

TABLE I STATBUNKER FOOTBALL STATISTICS

Player	Cristiano Ronaldo
Season	2014/2015
Position	Forward
Appearances	35
Right Foot	29
Header	12
Minutes played	3107
% Assists	14
Yellow Cards	5
Yellow - Away	3
Minutes per Yellow Cards	619.8
Sent Off	1
Minutes per Red Cards	3115
Goals	48

ข้อมูลชุดที่ 2 football events : เป็นสถิติของเหตุการณ์การทำประตูที่เกิดขึ้นในแต่ละการแข่งขัน เช่น นาทีที่ยิงประตู เท้าที่ใช้ยิง ความพยายามยิงประตูสำเร็จหรือไม่ อวัยวะที่ใช้ยิง จังหวะเหตุการณ์

TABLE II FOOTBALL EVENTS

player	lionel messi
time (minute)	17
text	Attempt saved. Lionel Messi (Barcelona) left footed shot from a difficult angle on the left is saved in the bottom left corner. Assisted by Pedro.
event_type	1 -attempt
side	1-home
shot_place	3-bottom left corner
shot_outcome	1-on target
location	7-difficult angle on the left-
bodypart	2-left foot
assist_method	1-pass
situation	1-open play
is_goal	0

IV. วิธีการทำนายผล

1. Linear Regression กับข้อมูลใน Table I. เพื่อทำนายจำนวนประตูที่ผู้เล่นยิงได้ จากตัวแปรต้นต่าง ๆ เช่น การใช้ศีรษะทำประตู การใช้เท้าประตูสามารถให้ผลการทำนายจำนวนประตูที่นักเตะยิงได้ โดย model ให้ค่า R-Square = 0.862 นอกจากนั้น model ให้ผลการคำนวณความสำคัญของตัวแปรต้นออกมาพบว่านักเตะที่เป็นกองหน้าซึ่งสามารถยิงประตูได้โดยใช้ศีรษะมีผลสำคัญกับจำนวนประตูได้ไปในทางเดียวกัน และจำนวนครั้งที่นักเตะส่งบอลให้เพื่อนทำประตูได้มีผลในทิศทางตรงกันข้ามกับจำนวนประตูที่ตนเองจะยิงได้

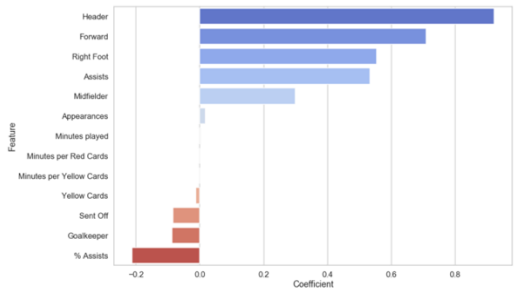


Fig.1.

2. Decision Tree กับข้อมูล Table II. เพื่อทำนายว่าผู้เล่นจะยิงเข้าประตูหรือไม่จากเหตุการณ์ที่เกิดขึ้นก่อนการยิงทำประตู ได้ผลลัพธ์เป็นตัวแปรที่มีระดับความสำคัญกับการยิงประตูเข้าดังแสดงใน Fig. 2 ซึ่งจะเห็นได้ว่าการยิงประตูตรงกรอบมีผลกับการได้ประตูสูงสุด รองลงมาคือ สถานการณ์ ณ ขณะนั้น) e.g. opened play, set piece, corner, free kick) และ ตำแหน่งการยิง โดยความแม่นยำของ model ดังแสดงใน Fig. 3

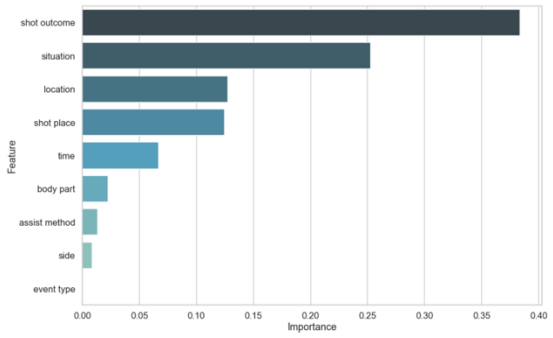


Fig.2.

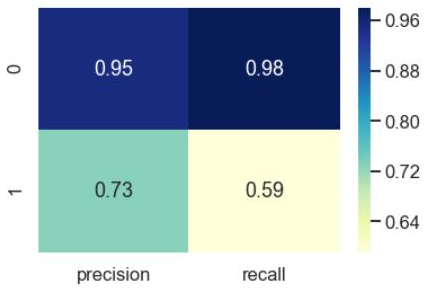


Fig.3.

V. สรุปผล

Linear regression model สามารถใช้ทำนายจำนวนประตูที่นักเตะจะทำได้ และ แสดงให้เห็นตัวแปรที่มีผลในทางส่งเสริมกับจำนวนประตูที่ยิงได้นั้น เช่น การใช้ศีรษะ ตำแหน่งกองหน้า ตัวแปรที่มีผลในทางตรงกันข้ามกับการยิงประตูได้เช่น อัตราการส่งบอลให้เพื่อนยิงประตู ทั้งนี้ตัวแปรที่ไม่มีผล เช่น จำนวนนาทีที่ลงเล่น ระยะเวลาต่อใบเหลืององที่ได้รับ ซึ่งจะสามารถนำไปใช้ประโยชน์ในวงการฟุตบอลเพื่อหานักเตะที่มีคุณสมบัติเหมาะสมสามารถทำประตูได้เยอะมาร่วมทีม

Decision tree สามารถทำนายว่านักเตะจะยิงประตูเข้าหรือไม่ภายในสถานการณ์ต่าง ๆ ตัวแปรสถานการร์ที่มีผลมากกับการยิงประตูได้ คือ การยิงตรงกรอบ การยิงในจังหวะ open play และตัวแปรที่มีผลน้อยเช่น side (การเล่นเป็นทีมเหย้าหรือเยือน (โดย modelสามารถให้คำทำนายที่มีค่าแม่นยำในระดับที่ยอมรับได้ สามารถนำไปใช้ให้เกิดประโญชน์ในการฝึกฝนนักเตะ วางแผนการเล่น เพื่อสร้างให้เกิดโอกาสที่จะยิงประตูได้สูงขึ้น

ACKNOWLEDGMENT

งานวิจัยฉบับนี้สำเร็จลุล่วงได้อย่างสมบูรณ์ด้วยความกรุณาอย่างยิ่งจาก ศ.ดร.ประภาส จงสถิตย์วัฒนา ที่ได้สละเวลาอันมีค่าแก่คณะผู้วิจัยให้คำปรึกษาและแนะนำ ตลอดจน ตรวจทานแก้ไขข้อบกพร่องต่าง ๆ ด้วยความเอาใจใส่เป็นอย่างยิ่ง จนงานวิจัยฉบับนี้สำเร็จสมบูรณ์ลุล่วงได้ด้วยดี นอกจากนี้ขอขอบคุณ คณาจารย์ ภาควิชาวิศวกรรมคอมพิวเตอร์ จุฬาลงกรณ์มหาวิยาลัย ที่ได้ถ่ายทอดวิชาความรู้ต่าง ๆ ที่นำมาปรับใช้ในการทำงานวิจัยฉบับนี้

REFERENCES

[1] Patrick Lucey, Alina Bialkowski, Mathew Monfort, Peter Carr and Iain Matthews;Disney Research.”Quality vs Quantity”: Improved Shot Prediction in Soccer using Strategic Features from Spatiotemporal Data.Sports Analytics Conference 2018

[2] Lotte Bransen,Pieter Robberechts,Jan Van Haaren,,Jesse Davis;SciSports;The Netherlands KU Leuven, Belgium.Choke or Shine?: Quantifying Soccer Players' Abilities to Perform Under Mental Pressure. Sports Analytics Conference 2019

[3] Javier Fernández , Luke Bornn , Dan Cervone.Decomposing the Immeasurable Sport: A deep learning expected possession value framework for soccer. Sports Analytics Conference 2019

[4] Eric P. Martin: Predicting Major League Baseball Strikeout Rates from Differences in Velocity and Movement Among Player Pitch Types.Sports Analytics Conference 2019

Aspect-Level sentiment analysis on movies review using Deep learning

Thananya Siriwatvanich, Chutimon Rungsilp, Thitima Tamoharanawong
Department of Computer Engineering
Faculty of Engineering
Chulalongkorn University
Bangkok, Thailand
thananya.sir@gmail.com, Chutimon.R@chula.ac.th, thitima.tamoS@gmail.com

***บทคัดย่อ*—ในวงการธุรกิจเพื่อเข้าใจความต้องการของผู้บริโภคเป็นเรื่องสำคัญอย่างมาก ดังนั้นการทราบถึงความคิดเห็นที่มีต่อสินค้าและบริการจึงเป็นสิ่งสำคัญ ในงานวิจัยนี้ได้นำการเรียนรู้แบบเชิงลึกมาสร้างแบบจำลองทำนายความรู้สึกจากชุดความคิดเห็นภาษาอังกฤษที่มีต่อภาพยนตร์เป็นไปในเชิงบวกหรือลบ โดยงานวิจัยนี้ได้เสนอการทำ CNN ร่วมกับ LSTM เข้าด้วยกันและทำการเปรียบเทียบแบบจำลองที่มีการนำเสนอในอดีต ได้แก่ Simple NN , CNN , LSTM และผลจากแบบจำลองที่ใช้ CNN และ LSTM ร่วมกันมีประสิทธิรูปที่สุดและค่าความแม่นยำ 86.69%**

***Abstract*—It is important to understand customers' needs. The customer opinion of products and services need to be considered. This work use deep learning to model the positive/negative feeling on movies review. We propose the use of convolutional neural network and long short term memory together and compare it with simple neural network. The result shows that the proposed model has the accuracy of 86.69%.**

Keywords— movie review, sentiment analysis, LSTM, CNN

I. ที่มาและความสำคัญ

เนื่องจากในปัจจุบันนี้อิทธิพลของโซเชียลมีเดียได้เข้ามามีส่วนสำคัญในการดำรงชีวิตของมนุษย์ โดยเฉพาะความคิดเห็นของผู้คนบนโซเชียลมีเดีย ที่มีผลอย่างมากกับการตัดสินใจของผู้บริโภคต่อธุรกิจหลายๆด้าน ไม่ว่าจะเป็นธุรกิจการท่องเที่ยว การศึกษา อุตสาหกรรมการบันเทิง การซื้อสินค้าและบริการต่างๆ โดยข้อความที่เป็นความคิดเห็นที่เกิดขึ้นนั้นมีทั้งในเชิงบวกและลบ ซึ่งผู้ผลิตหรือเจ้าของกิจการสามารถนำข้อมูลเหล่านี้มาใช้ในการพัฒนาสินค้าและบริการของตนเองได้

ในอดีตกระบวนการวิเคราะห์ความรู้สึก (Sentimental Analysis) นั้นจะใช้ Traditional Machine Learning เช่น Supporting Vector Machine (SVM), Naive Bayes, K-Nearest Neighbors (k-NN) แต่เนื่องจากประสิทธิภาพในการประมวลผลและความสามารถในการจัดการกับข้อมูลขนาดใหญ่ของคอมพิวเตอร์ในปัจจุบันเพิ่มสูงขึ้น ดังนั้นจึงมีความเป็นไปที่ได้จะประยุกต์ใช้เทคนิคการเรียนรู้ของเครื่องที่ซับซ้อนมากขึ้น โดยใช้แบบจำลองการเรียนรู้เชิงลึก (Deep Learning Model) เช่น Convolutional neural network (CNN), Recursive Neural Tensor Networks (RTNT), Multilayer Perceptron (MLP), Long short-term memory (LSTM) เป็นต้น เพื่อนำไปใช้ในการทำนายความรู้สึกได้แม่นยำมากยิ่งขึ้น

นอกจากนี้ในหลายๆงานวิจัยยังได้มีการนำเทคนิคของ Deep Learning Model มาใช้ร่วมกันในรูปแบบที่ซับซ้อนมากยิ่งขึ้น และผลจากการประยุกต์ใช้แบบจำลองเหล่านี้นั้นทำให้ความถูกต้องและความแม่นยำของการวิเคราะห์ความรู้สึกเพิ่มมากขึ้น

ดังนั้นในงานวิจัยนี้จึงได้มีการนำข้อมูลที่เป็นข้อความแสดงความคิดเห็นภาษาอังกฤษเกี่ยวกับภาพยนตร์มาวิเคราะห์และทำนายความรู้สึกของผู้แสดงความคิดเห็นว่าเป็นในเชิงบวกหรือลบ โดยการดึงคุณลักษณะที่ได้จากในแต่ละความคิดเห็นด้วยเทคนิคต่างๆ และนำคุณลักษณะดังกล่าวมาใช้ในการทำแบบจำลองการเรียนรู้เชิงลึกเพื่อทำนายประเภทของความคิดเห็นว่าเป็นเชิงบวกหรือลบ

II. งานวิจัยที่เกี่ยวข้อง

การศึกษาสำหรับการสร้างแบบจำลองการแยกประเภทความคิดเห็นในงานวิจัยต่างๆ ส่วนใหญ่จะประกอบกอบด้วย 2 ส่วนหลักที่ช่วยในการสร้างแบบจำลอง คือ การเลือกคุณลักษณะที่ใช้ในการวิเคราะห์อารมณ์ และการเลือกแบบจำลองที่ในการทำนายความรู้สึก ซึ่งได้กล่าวเพิ่มเติมต่อไปนี้

A. การเลือกคุณลักษณะที่ใช้ในการวิเคราะห์อารมณ์

งานวิจัยนี้ได้มีการเลือกคุณลักษณะจากข้อมูลความคิดเห็นมาใช้ในการทำแบบจำลองอารมณ์เชิงบวกและลบ โดยขั้นตอนแรกของการนำข้อมูลที่เป็นตัวอักษรมาใช้สร้างแบบจำลองเพื่อใช้ในการวิเคราะห์นั้น จะต้องมีการเตรียมข้อมูล เพื่อแปลงข้อความให้กลายเป็นตัวเลข และจะต้องมีขั้นตอนการคัดเลือกคุณสมบัติที่เหมาะสมก่อนนำมาใช้ในการทำแบบจำลองในการตัดสินใจ โดยเทคนิคในการคัดเลือกคุณสมบัติที่ได้จากข้อความมีด้วยกันหลายวิธี ได้แก่ Wise Tokenizer, Unigram และ Bigram [2 , 5] เป็นต้น ซึ่งเทคนิคดังที่กล่าวมาทำให้ได้เวกเตอร์คุณสมบัติ ที่จะนำไปใช้ในการทำแบบจำลองทำนายความรู้สึกได้ต่อไป

B. การเลือกแบบจำลองในการทำนายความรู้สึก

เพื่อให้ได้แบบจำลองที่มีประสิทธิรูปที่สุด จึงได้มีการนำเทคนิคต่างๆมาประยุกต์ใช้ร่วมกัน ไม่ว่าจะเป็น Naive Bayes, K-Nearest Neighbors (k-NN) [1], Supporting Vector Machine (SVM) [3,4], Recursive Neural Tensor Networks (RTNT), Multilayer Perceptron (MLP), Long short-term memory (LSTM), Convolutional neural network (CNN) [5] โดยผลจากการทำนายด้วยวิธีต่างๆ เราพบว่าเทคนิคการสร้างแบบจำลองเชิงลึก (Deep Learning) ให้ค่าเปอร์เซ็นต์ความถูกต้องแม่นยำที่ดีกว่าการใช้แบบจำลองทั่วไป

นอกจากนี้ เรายังสามารถนำข้อมูลความคิดเห็นในธุรกิจอื่นเข้ามาช่วยในการทำนายได้ โดยในงานวิจัยที่นำมาศึกษานั้น ได้มีการนำข้อมูลความคิดเห็นที่มีต่อการเข้าพักโรงแรม เข้ามาช่วยในการเรียนรู้แบบจำลอง โดยผลที่ได้จากการนำข้อมูลเหล่านั้นมาทำให้ได้ความถูกต้องที่มากขึ้นเช่นกัน

จากงานวิจัยที่กล่าวมาในข้างต้น จะเห็นว่ามีการสร้างแบบจำลองเพียงแค่ชั้นเดียวเท่านั้น ในงานวิจัยนี้จึงอยากจะพัฒนาโมเดลด้วยการสร้างแบบจำลองเชิงลึกหลายๆแบบมาประยุกต์ใช้ร่วมกัน

III. วิธีวิจัยและทฤษฎีบทที่เกี่ยวข้อง

ต่อจากนี้จะกล่าวถึงวิธีวิจัยและทฤษฎีบทที่เกี่ยวข้องในงานวิจัยนี้ โดยจะประกอบไปด้วย 4 ส่วน คือ ชุดข้อมูลที่ใช้ในการทำงานวิจัย การตัดคำและการเตรียมข้อมูล แบบจำลองและการประเมินผล และผลลัพธ์ โดยจะบรรยายต่อดังนี้

A. ชุดข้อมูล

ในงานวิจัยนี้ได้นำข้อมูลจาก www.kaggle.com ซึ่งเป็นเว็บไซต์ที่รวบรวมจากหลากหลายที่ โดยเป็นข้อมูลที่เปิดเผยต่อสาธารณะและไม่มีค่าใช้จ่าย โดยงานวิจัยนี้ได้นำข้อมูล IMDB Dataset of 50K Movie Reviews (ข้อมูลจำนวน 49,582 rows เก็บจาก IMDB ประกอบด้วย ความคิดเห็นภาพยนตร์ และ ความรู้สึกที่แบ่งเป็น 2 ประเภท ได้แก่ Positive และ Negative ดังตัวอย่างข้อมูลรีวิวภาพยนตร์ จาก IMDB Dataset ประกอบด้วยข้อมูล ส่วน คือ ความคิดเห็นและ ประเภทของความรู้สึก 2 แสดงในรูปที่ 1

	review	sentiment
0	One of the other reviewers has mentioned that ...	positive
1	A wonderful little production. The...	positive
2	I thought this was a wonderful way to spend ti...	positive
3	Basically there's a family where a little boy ...	negative
4	Petter Mattei's "Love in the Time of Money" is...	positive

รูปที่ 1 ตัวอย่างข้อมูล รีวิวภาพยนตร์ จาก IMDB Dataset

B. การตัดคำ และการเตรียมข้อมูล

เนื่องจากข้อมูลที่ได้จากการแสดงความคิดเห็นเป็นข้อมูลที่เป็นตัวอักษรซึ่งอยู่ในรูปของข้อความ จึงต้องมีการเตรียมข้อมูลก่อนนำไปสร้างแบบจำลอง โดยได้ทำการตัดประโยคและคำในประโยคออกจากกันด้วยการใช้ Word Tokenizer แล้วคัดกรองคำที่ไม่เกี่ยวข้องกับการตัดสินใจออก ได้แก่ ตัวเลข เครื่องหมาย และคำที่ไม่มีความหมาย (stop words) เช่น Article จากนั้นจึงทำการแปลงคำให้เป็นตัวเลขที่อยู่ในรูปแบบของเวกเตอร์ด้วยการใช้ Pre-train Model Word Embedding ของ GloVe (Global Vectors for Word Representation) ดังแสดงตัวอย่างข้อมูลใน รูปที่ 2-1 และรูปที่ 2-2 ดังนั้นในแต่ละความคิดเห็นจะได้เวกเตอร์ของตัวเลขที่แทนความคิดเห็นเพื่อนำไปใช้ในการสร้างแบบจำลองในขั้นตอนถัดไป


```
{'the': 1}{'and': 2}{'of': 3}{'to': 4}{'is': 5}
{'it': 6}{'in': 7}{'this': 8}{'that': 9}{'was': 10}
{'as': 11}{'movie': 12}{'for': 13}{'with': 14}{'but': 15}
{'film': 16}{'you': 17}{'on': 18}{'not': 19}{'are': 20}
{'he': 21}{'his': 22}{'have': 23}{'one': 24}{'be': 25}
{'all': 26}{'at': 27}{'they': 28}{'by': 29}{'an': 30}
{'who': 31}{'so': 32}{'from': 33}{'like': 34}{'there': 35}
{'or': 36}{'just': 37}{'her': 38}{'out': 39}{'about': 40}
```

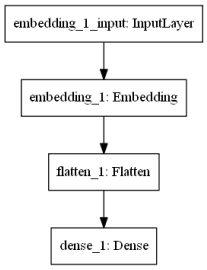

รูปที่ 2-1 ตัวอย่างดิกชันนารีของคำที่ได้จากข้อมูลในการ train


```
[[   1  296  140 ...  198  345 3812]
 [ 100   20  155 ...   82   99    6]
 [ 128 1307  108 ...    0    0    0]
 ...
 [ 778    8   24 ...    0    0    0]
 [   8  347   10 ...    0    0    0]
 [ 100  121   75 ...   67   69 1946]]
```

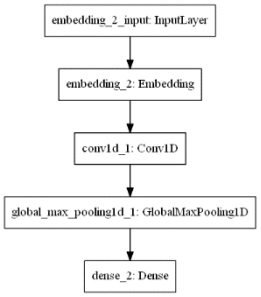

รูปที่ 2-2 เวกเตอร์ที่แสดงแต่ละข้อความ

C. แบบจำลองและการประเมินผล

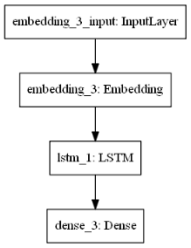
เพื่อให้สามารถประเมินประสิทธิภาพของข้อมูลได้ จึงทำการแบ่งชุดข้อมูลสำหรับสร้างแบบจำลอง (Training set) และชุดข้อมูลสำหรับทดสอบแบบจำลอง (Test set) เพื่อใช้ในการเปรียบเทียบแบบจำลอง เทคนิคการแบ่งข้อมูลเพื่อประเมินประสิทธิภาพดังที่กล่าวมาข้างต้นสามารถทำได้หลายวิธี และวิธีที่ได้รับความนิยมคือ เทคนิค k-fold cross validation ที่ช่วยลดการเกิด overfitting ของแบบจำลอง อันเป็นสาเหตุที่ทำให้แบบจำลองมีความเอนเอียงไปทางชุดข้อมูลที่มีลักษณะคล้ายกับชุดข้อมูลที่ใช้ในการสร้างแบบจำลอง หรือก็คือมีความถูกต้องสำหรับชุดข้อมูลเพียงบางกลุ่ม โดยในงานวิจัยนี้ได้มีการประเมินคุณภาพของแบบจำลองโดยการทำ 5 fold cross validation



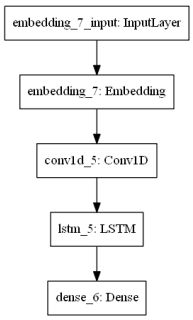
รูปที่ 3-1 พารามิเตอร์แบบจำลอง Simple NN



รูปที่ 3-2 พารามิเตอร์แบบจำลอง CNN



รูปที่ 3-3 พารามิเตอร์แบบจำลอง LSTM



รูปที่ 3-4 พารามิเตอร์แบบจำลอง CNN + LSTM

สำหรับงานวิจัยนี้ได้ประยุกต์ใช้ LSTM และ CNN มาใช้ร่วมกันโดยการใช้ CNN เพื่อสกัดจุดเด่นของเวกเตอร์ออกมาก่อน แล้วนำเข้าสู่โมเดล LSTM ในรูปที่ 3-4 เพื่อนำมาเปรียบกับแบบจำลองการเรียนรู้เชิงลึกพื้นฐานอื่นๆ ได้แก่ Simple NN, CNN และ LSTM ซึ่งพารามิเตอร์ที่ใช้ออกแบบแบบจำลองแต่ละแบบเป็นดังรูปที่ 3-1 ถึงรูปที่ 3-3

D. ผลลัพธ์

Model	**Accuracy**
Simple NN	73.93%
CNN	83.15%
LSTM	85.63%
CNN + LSTM	86.69%

ผลจากทดสอบพบว่าเมื่อนำแบบจำลอง Simple NN , CNN , LSTM และ CNN+LSTM ให้ค่าความแม่นยำ (Accuracy) เท่ากับ 73.93% , 83.15%, 85.63% และ 86.69% ตามลำดับ ซึ่งจะเห็นได้ว่า เมื่อมีการทำแบบจำลองที่ซับซ้อนมากขึ้น ทำให้ได้ค่าความแม่นยำสูงขึ้น 1.06% เมื่อเทียบกับงานวิจัยก่อนหน้า

IV. สรุปผล

ความคิดเห็นของผู้บริโภคเป็นส่วนสำคัญที่สะท้อนความต้องการที่แท้จริงของผู้บริโภค ดังนั้นการที่จะทราบได้ถึงความคิดเห็นต่อสินค้าบริการโดยรวมเป็นไปในทางบวกหรือลบจึงสำคัญ ในงานวิจัยในได้ทำการสร้างแบบจำลองเพื่อแยกความคิดเห็นในทางบวกและลบจากชุดข้อมูลแสดงความคิดเห็นภาพยนตร์ (IMDB Dataset of 50K Movie Reviews) โดยทำการเปรียบเทียบระหว่างแบบจำลองทั้ง 4 แบบ คือ Simple NN , CNN , LSTM และ CNN+LSTM และพบว่าแบบจำลองที่มีประสิทธิรูปที่สุดคือแบบจำลองที่รวมการใช้แบบจำลองระหว่าง CNN และ LSTM ซึ่งให้ความแม่นยำสูงสุดถึง 86.69%

เอกสารอ้างอิง

[1] Dey, L., et al. (2016). "Sentiment analysis of review datasets using naive bayes and k-nn classifier."

[2] Pang, B., et al. (2002). Thumbs up?: sentiment classification using machine learning techniques. Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10, Association for Computational Linguistics.

[3] Savoy, O. K. J. J. P. C. (2012). "Feature selection in sentiment analysis." 273-284.

[4] Schouten, K., et al. (2015). "Survey on aspect-level sentiment analysis." 28(3): 813-830.

[5] Singh, V. K., et al. (2013). Sentiment analysis of movie reviews: A new feature-based heuristic for aspect-level sentiment classification. 2013 International Mutli-Conference on Automation, Computing, Communication, Control and Compressed Sensing (iMac4s), IEEE.

Eye blink detection using machine learning for prevention of computer vision syndrome

Suppat Metarugcheep, Pimarn Kantithammakorn, Suthat Thanajan
Department of Computer Engineering
Faculty of Engineering
Chulalongkorn University
Bangkok, Thailand
{6270302421, 6270194021, 6270297921}@student.chula.ac.th

***บทคัดย่อ*—โรคคอมพิวเตอร์วิชั่นซินโดรมส่งผลกระทบต่อคนทำงานจำนวนมากที่อยู่กับคอมพิวเตอร์เป็นเวลานาน งานวิจัยนี้นำเสนอวิธีการตรวจจับการกระพริบตาเพื่อป้องกันโรคคอมพิวเตอร์วิชั่นซินโดรมด้วยการเรียนรู้ของเครื่องที่ให้ความแม่นยำสูง ในงานวิจัยนี้ได้นำเสนอวิธีการตรวจจับบริเวณรอบดวงตามาพิจารณาร่วมกับการเคลื่อนไหวของศีรษะเพื่อเพิ่มความแม่นยำในการตรวจจับการกระพริบตา ในขั้นตอนการตรวจจับการเคลื่อนไหวของศีรษะ เทคนิค Gradient Boosted Trees มีความแม่นยำถึง 100% และมีความเร็วในการประมวลผลมากที่สุดเมื่อเทียบกับการเรียนรู้ของเครื่องเทคนิคอื่น**

***Abstract* — Computer vision syndrome affects the users who spend long time with computer sceens. This work proposes a method to detect eye blinking to prevent it. The proposed method uses the detection of the area around eyes in combination with the motion of head to increase the accuracy to detect eye. To detect head motion, the gradient boosted tree has the accuracy to 100%. It is the fastest method compared to other methods.**

Keywords— computer vision syndrome, Gradient Boosted Trees, eye blink detection

I. INTRODUCTION

ปัจจุบันคอมพิวเตอร์เข้ามามีบทบาทอย่างมากในชีวิตประจำวันของมนุษย์ คอมพิวเตอร์กลายเป็นเครื่องมือหลักในการทำงานของสำนักงานทั่วโลก มีการประเมินว่าผู้คนนับล้านคนใช้เวลาทั่วโลกใช้เวลาอยู่กับหน้าจอคอมพิวเตอร์ โทรศัพท์มือถือ และอุปกรณ์อื่นที่มีหน้าจอแสงผลมากกว่า 3 ชั่วโมงต่อวัน งานวิจัยนี้จะกล่าวถึง และพยายามป้องกันปัญหาสุขรูปที่เกิดจากการใช้งานคอมพิวเตอร์ที่มีแนวโน้มสูงเพิ่มมากขึ้นโดยเฉพาะปัญหาเกี่ยวข้องกับสายตาที่เรียกว่า Computer Vision Syndrome (CVS)

CVS คือกลุ่มอาการทางตาที่เกิดจากการใช้คอมพิวเตอร์มากเกินไป เช่น ปวดตา ตาล้า ตาแดง ตามัว และตาแห้ง เป็นต้น การเพ่งสายตาหน้าจอคอมพิวเตอร์ต่อเนื่องเป็นเวลานานจะส่งผลให้อัตราการกะพริบตาลดน้อยลงกว่าปกติ ซึ่งทำให้ปริมาณน้ำตาที่ออกมาเคลือบตาลดลง เป็นผลให้ตาแห้ง แสบตา มองเห็นภาพไม่ชัด งานวิจัยของ Blehm et al. [1] แสดงให้เห็นว่ากว่า 70% ของผู้ใช้งานคอมพิวเตอร์ทั่วโลกประสบกับปัญหา CVS และพบจำนวนผู้ป่วยกำลังเพิ่มขึ้นอย่างน่ากังวล งานวิจัยของ Bali et al.[2] ระบุว่า CVS ส่งผลกระทบโดยตรงต่อประสิทธิภาพในการทำงานของผู้ป่วย จักษุแพทย์ และนักวิจัยชี้ว่าสาเหตุหลักของ CVS คือการใช้งานจอภาพคอมพิวเตอร์ที่เกินกว่า 3 ชั่วโมงโดยไม่มีการพักสายตา ตามผลงานวิจัยของ Akinbinu et al. [3]

งานวิจัยฉบับนี้นำเสนอวิธีการตรวจจับอัตราการกะพริบตาในระหว่างการใช้งานคอมพิวเตอร์ โดยประยุกต์ทฤษฎีการตรวจจับใบหน้า ทฤษฎีตรวจจับดวงตา และทฤษฎีการเรียนรู้ของเครื่อง มาเพิ่มความแม่นยำในการตรวจจับ เมื่อระบบพบว่าผู้ใช้งานคอมพิวเตอร์มีอัตรากระพริบตาต่ำกว่าค่ามาตรฐาน ระบบจะแจ้งเตือนไปยังผู้ใช้งาน เพื่อให้ผู้ใช้งานคอมพิวเตอร์รับรู้ถึงระดับของการกระพริบตาที่ต่ำลง และเปลี่ยนกิจกรรมไปทำกิจกรรมอื่นเป็นการพักสายตา ซึ่งจะช่วยป้องกัน และลดความเสี่ยงที่จะทำให้เกิดอาการตาแห้ง ซึ่งเป็นสาเหตุสำคัญที่จะก่อให้เกิดปัญหา Computer Vision Syndrome

II. RELATED WORK

วิธีการหลายแบบในงานวิจัยหลายถูกนำมาตรวจจับการกระพริบตาในมนุษย์ ซึ่งวิธีการแต่ละแบบใช้วิธีการ ทรัพยากร ให้และผลลัพธ์ไม่เท่ากันขึ้นอยู่กับหลายปัจจัย อาทิเช่น สภาพแวดล้อมของห้องทดลอง ความสามารถของอุปกรณ์จับภาพ ความสามารถของเครื่องคอมพิวเตอร์ และความแม่นยำไปจนถึงความเร็วของอัลกอริทึมที่ใช้ในการตรวจจับการกระพริบตา Divjak et al.[4] เสนอวิธีการประมาณการไหลของแสงด้วยการปรับสภาพการไกลควบคู่กับค่าเกณฑ์ที่ปรับเปลี่ยนได้ ผลของงานวิจัยแสดงให้เห็นว่าอัลกอริทึมมีความแม่นยาในการตรวจจับการกะพริบตามากกว่า 90% แต่วิธีการนี้ยังคงมีข้อผิดพลาด ถ้าหากตำแหน่งของดวงตามีการเคลื่อนไหวขึ้นลงที่รวดเร็วกะทันหัน Morris et al.[5] เสนอวิธีการตรวจจับการกะพริบตาด้วยวิธีการตรวจจับบริเวณรอบดวงตาควบคู่กับการใช้แผนที่ความแปรปรวน (Various Mapping) ซึ่งผลของงานวิจัยนี้ให้ความแม่นมากกว่า 95% แต่ยังคงพบข้อผิดพลาดหากศีรษะของผู้ใช้งานมีการเคลื่อนไหว T.N. Bhaskar et al.[6] เสนอวิธีการวัดความแตกต่างของคู่ภาพควบคู่กับ การประมาณการไหลของแสงด้วยทิศทางและขนาด ผลของงานวิจัยนี้ให้ความแม่นมากกว่า 97% แต่ยังคงมีความผิดพลาดหากมีการเคลื่อนไหวของศีรษะ

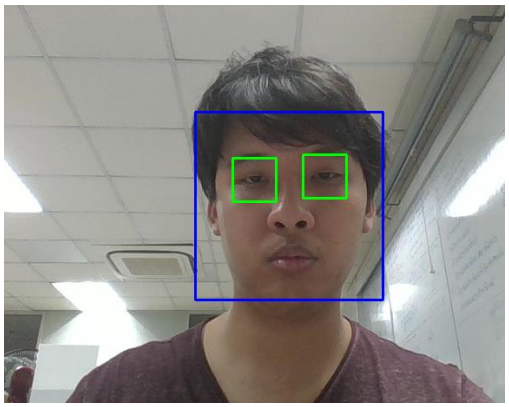
วิธีการดังที่กล่าวมาก่อนหน้านี้ ใช้การเคลื่อนไหวของบริเวณตาสำหรับการตรวจจับการกะพริบตา ซึ่งมีข้อเสียเปรียบตรงที่ เมื่อศีรษะมีการเคลื่อนไหว จะเกิดความผิดพลาดในการตรวจจับบริเวณตาที่ถูกต้อง Chau et al.[7] ใช้รูปภาพต้นแบบของการการเปิดตามาช่วยในการหาความสัมพันธ์และตรวจจับการกะพริบตาแบบเรียลไทม์ ซึ่งให้ผลแม่นยำสามารถรับข้อมูลรูปที่ได้จากกล้องวิดีโอของคอมพิวเตอร์ทั่วไป วิธีการดังกล่าวช่วยลดความซับซ้อนในการคำนวณแต่มีข้อจำกัดที่สามารถแยกสถานะของการเปิดหรือปิดตาอย่างชัดเจนเท่านั้น G. Pan et al.[8] นำเสนอแบบจำลองทางสถิติมาช่วยแยกแยะลักษณะของการปิดตาเพื่อช่วยทำให้การตรวจสอบการปิดตาแม่นยำขึ้นแม้ในกรณีที่ปิดตาไม่สนิทหรือมีการเคลื่อนไหวของบริเวณดวงตา

III. METHADOLOGY

ความแม่นยำในการตรวจจับการกระพริบตามีปัจจัยสำคัญที่ส่งผลต่อความแม่นยำคือการเคลื่อนที่ของศีรษะ ในเอกสารฉบับนี้ได้นำเสนอวิธีเพิ่มความแม่นยำในการตรวจจับการกระพริบตาด้วยการพิจารณาการเคลื่อนที่ของศีรษะเพื่อลดข้อผิดพลาดในการตรวจจับการกระพริบตา โดยการนำสองปัจจัยคือ ระยะห่างของเปลือกตาและการเคลื่อนที่ของศีรษะมาพิจารณาร่วมกัน

A. การตรวจจับการกระพริบตา

เพื่อลดขอบเขตในการตรวจจับลักษณะของดวงตา A. D. Joshi et.al.[9] นำเสนอการใช้เทคนิคการตรวจจับใบหน้าและดวงตามาประยุกต์ใช้กับรูปที่ได้จากกล้องก่อนเริ่มการตรวจจับการกระพริบตาโดยใช้กระบวนการขั้นตอนของ Viola และ Jones[10] ร่วมกับแบบจำลองสำหรับ haarcascades ของ Hameed[11] ดังรูปที่ 1 และนำผลที่ได้ไปใช้ในกระบวนการคำนวณหาความสูงของเปลือกตา



รูปที่ 1 การประยุกต์ใช้ กระบวนการขั้นตอนของ Viola และ Jones ร่วมกับแบบจำลองสำหรับ haarcascades ของ Hameed

จากการศึกษาของ M.H. Baccour et.al.[12] พบว่าในการกระพริบตาจะเกิดการเปลี่ยนแปลงความสูงของเปลือกตาในช่วง 0.4 วินาที ดังนั้น เอกสารฉบับนี้จึงใช้ลำดับของรูปที่ต่อเนื่องกัน 12 ภาพ มาเป็นหนึ่งในคุณลักษณะสำคัญของเครือข่ายประสาทเทียมเพื่อใช้ร่วมพิจารณากับการตรวจจับการเคลื่อนที่ของศีรษะ

B. การตรวจจับการเคลื่อนที่ของศีรษะ

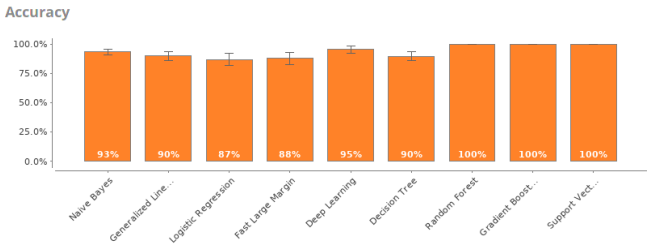
การพิจารณาการเคลื่อนที่ของศีรษะ งานวิจัยฉบับนี้ได้ประยุกต์ใช้งานวิจัยของ Zhao et al. [13] ที่มีการใช้เทคนิคการทำ optical flow เพื่อวิเคราะห์ลักษณะการเคลื่อนที่ของจมูกเพื่ออธิบายการเคลื่อนที่ของศีรษะ โดยงานวิจัยฉบับนี้ได้มุ่งเน้นที่การแยกลักษณะการเคลื่อนที่ของศีรษะตามปกติ กับการเคลื่อนที่ของศีรษะที่ผิดปกติออกจากกัน ซึ่งตามปกติแล้วศีรษะของคนเราที่กล้องจับได้จะไม่ได้อยู่ในตำแหน่งเดิมตลอดเวลาแต่จะมีการเปลี่ยนแปลงตำแหน่งเล็กน้อย

งานวิจัยฉบับนี้เสนอการใช้เทคนิคการเรียนรู้ของเครื่องในการแยกการเคลื่อนที่ของศีรษะตามปกติ และการเคลื่อนที่ของศีรษะที่ผิดปกติออกจากกัน โดยอาศัยลักษณะของการเคลื่อนที่ที่เกิดขึ้นในช่วง 0.4 วินาที เช่นเดียวกับการเปลี่ยนแปลงความสูงของเปลือกตา(ตามงานวิจัยของ M.H. Baccour et.al.[12]) เพื่อให้ผลลัพธ์ที่ได้จากทั้ง 2 ขั้นตอนสามารถพิจารณาร่วมกันได้อย่างตรงไปตรงมา โดยข้อมูลที่เก็บใน 1 แถวจะประกอบไปด้วย 25 คอลัมน์คือตำแหน่งของจมูกที่อ้างอิงจากใบหน้าในแนวตั้ง และ ตำแหน่งของจมูกในแนวนอน จำนวน 12 ชุด และป้ายกำกับที่ระบุว่าข้อมูลจำนวน 12 ชุดข้างต้น มีการเคลื่อนที่ในลักษณะปกติหรือไม่ โดยชุดข้อมูลที่มีการเคลื่อนที่ในลักษณะปกติ จะมีป้ายกำกับว่า non_move และชุดข้อมูลที่มีลักษณะการเคลื่อนที่ที่ผิดปกติ จะมีป้ายกำกับว่า move ตามรูปที่ 2

X1	Y1	X2	Y2	X3	Y3	X4	Y4	X5	Y5	X6	Y6	X7	Y7	X8	Y8	X9	Y9	X10	Y10	X11	Y11	X12	Y12	label
340	256	341	257	341	258	341	257	341	257	362	306	341	257	341	258	341	258	341	257	341	258	341	257	non_move
341	257	341	258	341	257	341	257	362	306	341	257	341	258	341	258	341	257	341	258	341	257	341	258	non_move
341	258	341	257	341	257	362	306	341	257	341	258	341	258	341	257	341	258	341	257	341	258	341	258	non_move
341	257	341	257	362	306	341	257	341	258	341	258	341	257	341	258	341	257	341	258	341	258	341	258	non_move
341	257	362	306	341	257	341	258	341	258	341	257	341	258	341	257	341	258	341	258	341	258	342	258	non_move
362	306	341	257	341	258	341	258	341	257	341	258	341	257	341	258	341	258	341	258	342	258	341	256	non_move
353	261	357	258	360	260	364	260	370	262	380	264	380	264	366	262	346	262	327	261	312	262	311	263	move
357	258	360	260	364	260	370	262	380	264	380	264	366	262	346	262	327	261	312	262	311	263	322	262	move
360	260	364	260	370	262	380	264	380	264	366	262	346	262	327	261	312	262	311	263	322	262	348	263	move
364	260	370	262	380	264	380	264	366	262	346	262	327	261	312	262	311	263	322	262	348	263	374	266	move
370	262	380	264	380	264	366	262	346	262	327	261	312	262	311	263	322	262	348	263	374	266	387	270	move
380	264	380	264	366	262	346	262	327	261	312	262	311	263	322	262	348	263	374	266	387	270	374	277	move

รูปที่ 2 ตัวอย่างข้อมูล

จากชุดข้อมูลตัวอย่างที่เก็บมาทั้งหมด ได้แบ่งออกเป็นชุดข้อมูลสำหรับฝึกแบบจำลองและชุดข้อมูลสำหรับทดสอบวัดประสิทธิภาพแบบจำลอง ในแบบจำลองการเรียนรู้ของเครื่องชนิดแยกประเภทข้อมูล (classification machine learning model) ผู้วิจัยพบว่าแบบจำลอง Random forest, Gradient Boosted Trees และ Support Vector Machine มีความแม่นยำถึง 100% ตามรูปที่ 3 และเมื่อพิจารณาถึงความเร็วในการนำแบบจำลองไปใช้งานจริง จะพบว่าแบบจำลอง Gradient Boosted Trees ใช้ระยะเวลาน้อยที่สุดในขั้นตอนของการทดสอบวัดประสิทธิภาพแบบจำลองตามตารางที่ I



รูปที่ 3 ความแม่นยำของแบบจำลอง

ตารางที่ I ระยะเวลาที่ใช้ในการประมวลผล

Model	Training time (ms)	Testing time (ms)
Naive Bayes	749.34	933.34
Generalized Linear Model	1098.67	726.67
Logistic Regression	1120.0	353.34
Fast Large Margin	661.34	593.34
Deep Learning	2474.67	686.67
Decision Tree	325.34	293.34
Random Forest	1386.67	2853.34
Gradient Boosted Trees	7437.34	426.67
Support Vector Machine	296.0	3126.67

IV. CONCLUSION AND FUTURE WORK

การตรวจจับการกระพริบตาและการหาทิศทางการเคลื่อนที่ของจมูกยังมีความท้าทายและข้อจำกัดในด้านความแม่นยำในการตรวจหาคุณลักษณะของใบหน้าจากภาพ ซึ่งแบบจำลองสำหรับ haarcascades ของ Hameed จะเน้นการตรวจจับใบหน้าในลักษณะหน้าตรง ทำให้วิธีการที่เสนอในงานวิจัยฉบับนี้ไม่สามารถใช้งานได้ดีเท่าที่ควร อย่างไรก็ตามข้อจำกัดนี้

สามารถแก้ไขได้โดยการสร้างแบบจำลองสำหรับ haarcascades ใหม่ โดยใช้ชุดข้อมูลที่ประกอบไปด้วยข้อมูลลักษณะใบหน้าทั้งหน้าตรงและไม่ใช่หน้าตรง ในงานวิจัยฉบับนี้ยังมีการเก็บข้อมูลที่น้อยเกินไปทำให้ไม่สามารถทดสอบและแยกประสิทธิภาพของแบบจำลอง Random forest, Gradient Boosted Trees และ Support Vector Machine ออกจากกันได้อย่างชัดเจน ผู้วิจัยคาดหวังว่าถ้ามีการเก็บข้อมูลมากขึ้นแบบจำลองทั้ง 3 ชนิดจะสามารถวัดประสิทธิภาพได้ดียิ่งขึ้น

REFERENCES

[1] C. Blehm, S. Vishnu, S. Mitra, A. Khattak, Yee RW , “Computer Vision Syndrome: A Review,” Survey of Ophthalmology, Elsevier , vol. 50, no. 3, 2005.

[2] Jatinder Bali, Navin Neeraj, Renu Thakur Bali,“Computer Vision Syndrome: A Review,” Journal of clinical ophthalmology and research, vol. 2, pp. 61-68, 2014.

[3] Akinbinu T.R., Mashalla Y.J., “Impact of computer technology on health:Computer Vision Syndrome,” Medical Practice and Review, vol. 5, pp. 20-30, 2014.

[4] Divjak M, Bischof H. , “Eye blink based fatigue detection for prevention of Computer Vision Syndrome,” MVA 2009 IAPR conference on machine vision applications, 2009.

[5] T.Morris,P. Blenkhorn,F. Zaidi,“Blink detection for real-time eye tracking,” Journal of Network and Computer Applications, vol. 25, 129-143.

[6] T.N. Bhaskar, Foo Tun Keat, S. Rangnath, Y.V. Venkatesh, “ Blink detection and eye tracking for eye localization,” TENCON 2003. Conference on Convergent Technologies for the Asia- Pacific region, 2003.

[7] W. Chau, A.R. McIntosh, S.E. Robinson, M. Schulz, C. Pantev “Improving permutation test power for group analysis of spatially filtered MEG data” NeuroImage, 23 (3) (2004), pp. 983-996.

[8] G. Pan, L. Sun, Z. Wu, S. Lao, "Eyeblink-based anti-spoofing in face recognition from a generic webcamera", Proc. IEEE 11th Int. Conf. Comput. Vis. (ICCV), pp. 1-8, Oct. 2007.

[9] A. D. Joshi, A. A. Kadethankar, V. P. Patwardhan, "Eye blinking detection for the detection of computer vision syndrome", 2017 Innovations in Power and Advanced Computing Technologies (i-PACT), pp. 1-3, 2017.

[10] P. Viola, M. Jones, "Rapid object detection using a boosted cascade of simple features", Conference on Computer Vision and Pattern Reognition, 2001.

[11] Shameem Hameed “ haarcascade_eye” [online] . Available :https://github.com/opencv/opencv/blob/master/data/haarcascades.

[12] M. H. Baccour, F. Driewer, E. Kasneci and W. Rosenstiel, " Camera- Based Eye Blink Detection Algorithm for Assessing Driver Drowsiness," 2019 IEEE Intelligent Vehicles Symposium (IV), Paris, France, 2019, pp. 987-993.

[13] M. H. Baccour, F. Driewer, E. Kasneci and W. Rosenstiel, " Camera- Based Eye Blink Detection Algorithm for Assessing Driver Drowsiness," 2019 IEEE Intelligent Vehicles Symposium (IV), Paris, France, 2019, pp. 987-993.

Detecting a sign of Major Depressive Disorder from social network activities

Thanaphat Patravinij, Arunee Sridee, Natchapol Thongruang
Department of Computer Engineering
Faculty of Engineering
Chulalongkorn University
Bangkok, Thailand
{6270116021, 6270314021, 6270074321}@student.chula.ac.th

***Abstract*—Major depressive disorder (MDD) is becoming one of the most common mental illness today. According to Word Health Organization report in 2001, the number of patients from MDD is growing rapidly. Not all the patients have received adequate treatment therefore. We have received information from CLEF 2017 meeting where they gathered data in Reddit post from MDD patient. Based on these data, we will find the most accurate feature to early detect person who might be possible to have MDD.**

Keywords—depression, mental illness, MDD

I. INTRODUCTION

Social network is a very worldwide activity in our daily life. You may notice that most of the people spend time on their telephone during on public transporting almost every day and every time. Most of the information in the social network is very accurately because people think that social network is one of their comfortable area to express their feeling, record their daily activity (same as diary). Some people might use social network to contact with their friends by personal message or via friend's profile. These behaviors give us an opportunity to analyze their information which might lead to depressive disorder.

Normally when people are sick, there are several signs or symptoms that clearly shown and could be used to diagnose right away in this modern world with these advancement in technology and medical knowledge. However, there are some sicknesses that are harder to detect, compare to the obvious ones. Some of them are the hardest to be detected and diagnosed as the symptoms are not obviously shown or not be able to notice it at all. Mental illness is among the hardest group and even harder to know since the symptoms are not physically shown out. People with mental illness must notice their own sickness and be self-aware enough to know that they have a mental illness problem or else they wouldn't know and wouldn't even have a chance to receive a proper treatment.

Mental illness is now a leading cause of disability around the world. Major Depressive Disorder (MDD or simply depression) is one of the most common mental illness that largely spread around the world. According to World Health Organization, more than 300 million people of all ages suffer from depression globally. It can cause the affected person to suffer greatly and function poorly at work, at school and in the family. At its worst, depression can lead to suicide. Close to 800,000 people die due to suicide every year. Suicide is the second leading cause of death in 15-29 year olds. Although there are known, effective treatments for depression, fewer than half of those affected in the world (in many countries, fewer than 10%) receive such treatments. In countries of all income levels, people who are depressed are often not correctly diagnosed, and others who do not have the disorder are too often misdiagnosed and prescribed antidepressants. The fact and statistics about depression is quite terrifying. That is why we want to get involved in this area, to help the people with things we do best.

Social media is undeniably the most commonly used platform worldwide leading by the well-known brands like Facebook, Twitter, Instagram, etc. It is the place where most people with internet access spend their time on. The new generation people even have their social media profiles created before they even born. On social media, people tend to show their opinion and thoughts easier than it is in real life. They can even share their experience through text and storytelling, photos, videos, or even live videos which will be streamed from their device in real-time. It is true that not everyone has a Facebook account, but people usually have at least 1 type of social media account not necessary be Facebook. The popularity of social media helps people generate a huge amount of individual data which is purely focused on their personality with some of them didn't even realize it.

The inaccurate assessment and lack of tools to properly diagnose depression is the gap that we would like to fill-in here. Our goal is not to accurately detect those who suffer from depression so perfectly that no one would have depression anymore. The problem that we acknowledge here is that, it is hard to reach the people that doesn't regularly observing their selves. We want to increase the reach of depression detection from just using medical data to analyzing personal data that people generated. That way, we can evaluate their risk of having depression earlier before they need any medical treatment. From the social media activities, we believe that we could determine the risk of having a depression and identify the potential patients before they would actually have a depression. Obviously, this method would not work for people that doesn't have their personal data created on social media. But we believe that, the number of social media users will be growing in the future as well as the number of platforms available.

II. RELATED STUDIES

Several previous studies have highlighted the importance of early detection in improving outcomes related to MDD. Cacheda, Fernandez, Novoa[5] observed significant improve model of early detection of Depression by use 2 random forest model in Machine Learning technique, 1 RF use to detect depressed and 2nd independent RF to detect non-depressed base on textual, semantic and writing similarities. Choudhury, Gamon, Counts [1] study use data set from crowdsourcing to detect depression behavior, Engagement, emotion, depression language, by SVM technique. An SVM classifier can

predict ahead of the reported onset of depression of an individual with high accuracy. Stankevich, Isakov, Devyatkin and Smirnov [2] use CLEF/eRisk dataset to create classification model and compare the feature in the research to find out suitable feature to predict MDD, build the prediction model and use feature to learn. Separate feature to 3 groups, TF-IDF, World Embedding and Bi-Grams. On each group analyze compare between only feature, feature with Morphology, feature with Stylometric or feature with Morphology and Stylometric.

On our study we will focus dataset from reddit and study from various Machine learning technique to build a model to detect between depression and nondepression.

III. DATA

The research area of this paper will be focusing on the dataset from the CLEF/eRisk 2017 (Conference and Labs of the Evaluation Forum: early risk prediction on the Internet). The CLEF/eRisk 2017 was the first year that they held this conference. There were two possible ways to participate which are Research Papers and Pilot Task. The pilot task is an exploratory task on a topic of "Early risk detection of Depression". The challenge consists of sequentially processing pieces of evidence and detect early traces of depression as soon as possible. The task is mainly concerned about evaluating Text Mining solutions and, thus, it concentrates on texts written in Social Media.

The test collection for this pilot task is the collection described in [6]. It is a collection of writings (posts or comments) from a set of Social Media users. The collection contains textual interactions (posts or comments) from 892 users. 137 subjects have explicitly declared that they have been diagnosed with depression, and the remaining 755 subjects are a control group. For each subject, a (usually long) history of writings (posts or comments from a social networking site) is available. This is stored as an XML file (one per subject) with the following structure:

ID: contains the anonymised id of the subject.

TITLE: title of the post if available (if it is a comment then TITLE is empty)

INFO: additional info about the writing (source of the post/comment)

TEXT: body of the post or comment

```
<INDIVIDUAL>
<ID> ... </ID>
<WRITING>
<TITLE> ...   </TITLE>
<DATE> ... </DATE>
<INFO> ... </INFO>
<TEXT> ...  </TEXT>
</WRITING>
<WRITING>
<TITLE> ... </TITLE>
<DATE> ... </DATE>
<INFO> ... </INFO>
<TEXT> ... </TEXT>
</WRITING>
....
</INDIVIDUAL>
```

Fig. 1. Data structure

IV. EVALUATING DEPERESSIVE FEATURES

The data from [6] is stored as plain text XML format. We would like to gain some insight from these raw data before we could proceed and use them in our research. The data is structured in a collection of each user's posts and comments history records. Each individual will have an ID and a bunch of WRITING history data as their attributes. The structure of WRITING data contains TITLE, DATE, INFO and TEXT.

Creating the predictive model, between 2 classes of one being a depression patient and another is not, will be done by using some refined features that we could extract from the given dataset above. We could extract various information from the data but not all of them are useful enough to give us a precise and accurate result for our predictive model. In this work, our predictive model will be trained, using features like user engagement, user emotion, textual spreading and timespan.

A. *Engagement*

1. Volume: Define as the normalized number of posts per day made by the user.
2. Reply: Posts reply from a user per day indicating her level of social interaction with other Twitter users.
3. Links: Proportion of links (urls) shared by each user over a day.
4. Insomnia index: Showing depression signs tend to be active during evening and night. Hence we define a "night" window as "9PM-6AM" and define "day" window as"6.01AM-8.59PM" (Insomnia index=difference between "night" and "day" post)

We can apply dataset from reddit which is already classified from person who has depressive disorder.

B. *Emotion*

For this feature we can use reference of LIWC software.

Example. I feel so sad to day because I got argued with my girlfriend and we have a little fight but at the end everything went well

TRADITIONAL LIWC DIMENSION	YOUR DATA	AVERAGE FOR PERSONAL WRITING
I-WORDS (I, ME, MY)	11.5	8.70
SOCIAL WORDS	19.2	8.69
POSITIVE EMOTIONS	3.8	2.57
NEGATIVE EMOTIONS	11.5	2.12
COGNITIVE PROCESSES	15.4	12.52
SUMMARY VARIABLES		
ANALYTIC	32.3	44.88
CLOUT	77.9	37.02
AUTHENTICITY	99.0	76.01
EMOTIONAL TONE	1.0	38.60

Fig. 2. LIWC structure

Finally we can classify that our message is tended to be positive or negative meaning

C. *Texual Spreading*

We characterized the textual spreading of the writings by measuring the number of words used in each of the writings. Textual spreading measures the amount of textual information provided by the subject in their writings, and collected data in the following features:

- AvgWords: The average number of words per writing.
- DevWords: SD for the number of words per writing.
- MinWords: Minimum number of words in the subject's writings.
- MaxWords: Maximum number of words in the subject's writings.
- MedWords: Median for the number of words in the subject's writing.

D. *Time Span*

Another group of features was used to profile the moment when the writings were created. This information was expected to model differences in behavior among subjects diagnosed with depression versus non-depression. The following time features were proposed:

- Day: Percentage of writings provided by the subject, for each day of the week.
- Weekday: Accumulative percentage for all writings created in a weekday.
- Weekend: Accumulative percentage for all writings posted during the weekend.
- Hour: The hours of the day are divided into 4 homogeneous classes (0:00-5:59, 6:00-11:59, 12:00-17:59, and 18:00-23:59) and the percentage of writings that fall into each class is calculated.

As a summary, textual and semantic features are computed and aggregated for each user in comparison with all other users (grouped as positive and negative), meanwhile WFs are independently calculated and aggregated for each individual with respect to their postings.

V. PREDICTING DEPRESSIVE BEHAVIOR

We will now be using supervised learning technique to construct classifiers trained to predict depression in our two user classes. To avoid overfitting, we employ principal component analysis (PCA), although we report results for both all dimension-inclusive and dimension-reduced cases. We compare several different parametric and non-parametric binary classifiers to empirically determine the best suitable classification technique. The best performing classifier was found to be a Support Vector Machine classifier. For all of our analyses, we use cross validation on the set of 476 users.

VI. RESULT

We now focus on prediction of future episodes of depression. We first present some results of statistical significance of the behavioral features, as measured through their mean and variance values over the ground truth data analysis (Table I).

TABLE I. STATISTICAL SIGNIFICANCE COMPARING THE DEPRESSION AND NON-DEPRESSION CLASSES.

	Mean	Variance
Volume	15.21	14.88
Reply	22.88	13.89
Links	8.205	7.14
Insomnia Index	7.29	6.91

We present the results of these prediction models here. The results indicate that the best performing model (dimension-reduced features) in our test set yields an average accuracy of ~60% and high precision of 0.61, corresponding to the depression class. Next, we note that, one of the main characteristics of depression is disturbed cognitive processing of information, as well as a reduced sense of interest or motivation in day-to-day activities. Hence we observe better performance of depression language features in the prediction task. Finally, we conclude that social media activity provides useful signals that can be utilized to classify and predict whether an individual is likely to suffer from depression in the future.

VII. CONCLUSION

We have demonstrated the potential of using Reddit posting information as a tool for measuring and predicting major depression in individuals. First we asked for a collected and labelled data to receive the most useful information of depression, and proposed a variety of social media measures such as language, emotion, textual spreading and timespan to characterize depressive behavior. Our findings showed that individuals with depression show lowered social activity and greater negative emotion. Finally, we leveraged these distinguishing attributes to build an SVM classifier that can predict, ahead of the reported onset of depression of an individual, his/her likelihood of depression. The classifier yielded promising results with 60% classification accuracy. We hope to understand how analysis of social media behavior can lead to the development of scalable methods for automated public health tracking at-scale. We are also interested in harnessing the potential of social media in tracking the diffusion of affective disorders in populations in a nuanced manner; for identifying the incidence and impact of trauma on individuals during crisis events, and for modeling of help-seeking behavior, health risk behaviors, and risk of suicide.

REFERENCES

[1] De Choudhury, M., Gamon, M., Counts, S., & Horvitz, E. "Predicting Depression via Social Media". In Proc ICWSM-13, 2013.

[2] Stankevich, M., Isakov, V., Devyatkin, D., Smirnov, I. "Feature engineering for depression detection in social media". In ICPRAM, pp. 426–431, 2018.

[3] Md. Rafiqul Islam, Muhammad Ashad Kabir, Ashir Ahmed, Abu Raihan M. Kamal, Hua Wang and Anwaar Ulhaq. “Depression detection from social network data using machine learning techniques”. 2018.

[4] Nazanin Andalibi, Pinar Ozturk and Andrea Forte. “Depression-related Imagery on Instagram”. 2015.

[5] Fidel Cacheda, Diego Fernandez, Francisco J Novoa, Victor Carneiro. “Early Detection of Depression: Social Network Analysis and Random Forest Techniques”. 2019.

[6] David Losada, Fabio Crestani. “A Test Collection for Research on Depression and Language use”. In Experimental IR Meets Multilinguality, Multimodality, and Interaction 7th International Conference of the CLEF Association, CLEF 2016, Évora, Portugal, September 5-8, 2016