Machine Learning Algorithms for Natural Language Processing Tasks: A Case of COVID-19 Twitter data (Thailand)

[1]KUNYANUTH KULARBPHETTONG, [1]RUJIJAN VICHIVANIVES, [2]PANNAWAT KANJANAPRAKARN, [2]KANYARAT BUSSABAN, [2]JARUWAN CHUTRTONG, [3]NAREENART RUKSUNTORN

[1]Computer Science Program Suan Sunandha Rajabhat University Bangkok, THAILAND
[2]Faculty of Science and Technology Suan Sunandha Rajabhat University Bangkok, THAILAND
[3]Robotics Engineering program Faculty of Industrial Technology Suan Sunandha Rajabhat University Bangkok, THAILAND

Abstract: This paper presents the use of natural language processing for the problem of information extraction and sentiment analysis. The dataset is from Twitter that has the information of people mentioning about COVID-19, this study has two tasks: (i) classification approach for information extraction task and (ii) deep learning approach for sentiment analysis task. In information extraction task, the data was gathered from twitter that related to COVID-19 information, and the sequence labelling method applied to classify text before giving it to classification algorithms (K-NN, Naïve Bayes, Decision Tree, Random Forest, and SVM). In sentiment analysis task, data was classified by convert the word into index and using word embedding, then to process deep learning algorithm (Bi-directional GRU). The accuracy of two tasks are 98% and 79% respectively.

Keywords: COVID-19, KNN,Deep learning, Random Forest, Bi-directional GRU

Received: March 24, 2022. Revised: October 18, 2022. Accepted: November 21, 2022. Published: December 31, 2022.

1. Introduction

The Covid-19 epidemic situation reflected significant changes in many dimensions, both in human behavior consumption and service behavior. The SARS-Cov-2 known as Covid-19 has spread around the world and it's also designation as a worldwide pandemic by the World Health Organization in March 2020 [1]. While the world is struggling to handle with Covid-19, the number of infected patients has continuingly increased. Thailand has changed since Covid-19 and people are worried about the spread of this situation. Information is important to be aware of the events in society. People tend to follow the news through various media, including real news, fake news, or current news. Especially in this era of COVID-19, information has a huge impact on people's emotions. Therefore, most people have anxiety, paranoia, and fear that they or close people are infected which results in confusion in information. Twitter is one of the popular social media platforms that aims to provide users with the ability to comment in short texts up to 140 characters. Twitter presents the trend that people talk about or what trending is on Twitter right now. On twitter message, a hashtag is used to play critical roles in recent social movements such as #election, #Covid-19, and etc. It is a word or sentence that has a "#" preceding it. This is a form of metadata tag that is widely used in social media. Hashtags have played the important role in conversation. There was a discussion to exchange comments and hashtags in Thailand are used in a variety of ways.

Opinion mining is the science of gathering opinions from multiple messages on a particular subject to analyze opinions. It is often analyzed as positive, negative, or neutral. Information extraction and sentiment analysis has been broadly acknowledged as one of the first stages in the natural language processing [2], [3]. This research is aimed to classify the textual information on the social media platforms like Twitter. The significant approaches, like K-NN, Naïve Bayes, Decision Tree, Random Forest and Support Vector Machine (SVM), were used to information extraction process and then evaluated with the F1 score accuracy of each algorithm. In the second process, Bi-directional GRU, one of the deep learning method, was applied to use for sentiment analysis task. Experimental result may be in charge of helping to the development of public health properly.

2. Literature Reviews

This section presented a literature review of relevant researches for exploration an overview of current knowledge of sentiment analysis. Tweets from twitter [3] were classified into positive, negative, and neutral. Dusmanu et al. [4] applied argument mining methods to classify arguments on Twitter from actual facts. Vaccine-related tweets were analyzed and the results showed the number and the opinion polarity of tweets in neutral 60%, 23% against vaccination, and 17% in favor of vaccination [5].

Naïve Bayes model was implemented to analyze sentiments towards COVID-19 with Twitter datasets in English and Filipino language and the algorithm supports to classify tweets by using Rapid Miner [6]. Machine learning algorithms and lexicon-based approaches were proposed to sentiment word detection and POS tagging [7]. According to Tang, Kay and He [8], Naive Bayes (NB), and Support Vector Machine (SVM) were used to Text Classification. To analyze reliability, Naïve Bayes was adopted to identify the untrusted content on Twitter [9]. Deep learning based on

LSTM, GRU, and CNN and feature-based methods were combined to financial sentiment analysis [10]. Contextual deep learning was applied to analyze in sentiment analysis involves categorizing subjective opinions from text, audio, and video sources [11].

3. Methodology

This section describes the relevant approaches using conduct this research.

3.1 Data Set and Data Preparation

The scope of this study is Thailand and the data was considered news about COVID-19 in Thailand. In figure 1, data was collected almost 600,000 tweets by using Tweepy (a python library) [12] and then processed the raw data (Unstructured data) to be data that is in the form of an appropriate structure (structure data). The data was pre-processed by cleaning and tokenization text using NLTK library.

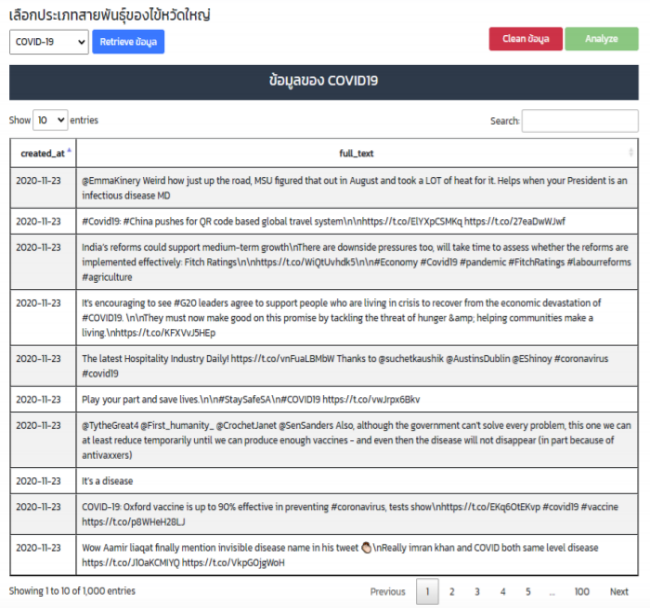

Fig. 1. A sample of data retrieved from the web page application

3.2 Information Extraction

The task consists in classifying a tweets as containing report information of coronavirus focused on particular tweets patterns like "total 42 cases" or "500 total deaths" of their sources. The five algorithms of classification including K-NN, Naïve Bays, Decision Tree, Random Forest and Support Vector Machine were used to classify tweets and evaluate the results.

The labelling is the next process from pre-processing and 700 tweets were selected to find the amount of people who affected in this pandemic. Then the data was annotated the numbers that occur in text. This would allow us to understand and make it easier to train the algorithms. These numbers are annotated as "1" if it follows by "total cases", or we annotated as "2" if it follows by "total deaths", if it not fits the above conditions then we annotated as "0" (see example (a) and (b) below).

(a) Text: "iran reports 3 new cases bringing total confirmed cases 52 total deaths."

Tag: "[0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 2, 0, 0]"

(b) Text: "total deaths 75 total cases 100."

Tag: "[0, 0, 2, 0, 0, 1]"

After manual labelling, the data set was prepared to be processed with algorithms by transform "Text" to Python dictionary by having key as the following.

"word" : The word itself.

"postag" : The part of speech of the word.

"nextword" : The next word after the word itself.

"nextwordtag" : The next part of speech tag of next word.

"previousword" : The previous word before the word itself.

"previoustag" : The previous part of speech tag before the word itself.

The data was spited it to 70:30 proportion training and testing and transform Python dictionary to vector by using DictVectorizer from Scikit-learn library [13].

4. Results

This section presented the results of this research. K-NN, Naïve Bays, Decision Tree, Random Forest and Support Vector Machine were used to extract information and the results were shown in table I and figure 2and 3.

TABLE I. RESULTS FOR INFORMATION EXTRACTION TASK

	Precision	Recall	F1
K-NN	0.93	0.91	0.92
Naïve Bayes	0.87	0.87	0.87
Decision Tree	0.93	0.90	0.92
Random Forest	0.95	0.91	0.93
SVM	0.94	0.91	0.93

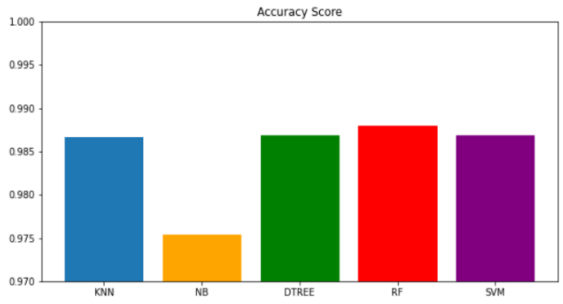

Fig. 2. Results of Accuracy Scores

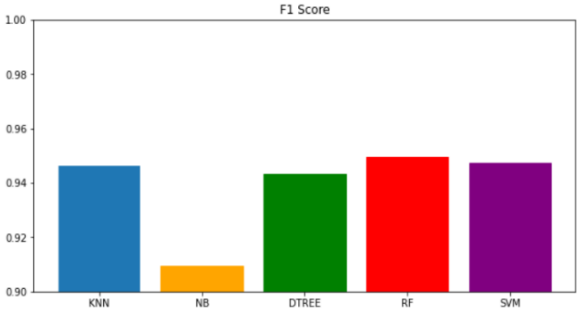

Fig. 3. Results of F1 Scores

When considered in each algorithm, the results were as follows:

(a) KNN has an overall accuracy of 98%, a classification precision in infected cases (1) 94% and precision in classifying deaths (2) 93% as shown in figure 4.

```
Accuracy: 0.9866387215090385
              precision    recall  f1-score

           1       0.94      0.95      0.94
           2       0.93      0.87      0.90
```

Fig. 4. Results of KNN

(b) Naïve Bays has an overall accuracy of 97%, a classification precision in infected cases (1) 85% and precision in classifying deaths (2) 90% as shown in figure 5.

```
Accuracy: 0.9753733298401887
              precision    recall  f1-score

           1       0.85      0.91      0.88
           2       0.90      0.82      0.86
```

Fig. 5. Results of Naïve Bays

(c) Decision Tree has an overall accuracy of 98%, a classification precision in infected cases (1) 94% and precision in classifying deaths (2) 93% as shown in figure 6.

```
Accuracy: 0.9866387215090385
              precision    recall  f1-score

           1       0.94      0.95      0.94
           2       0.93      0.87      0.90
```

Fig. 6. Results of Decision Tree

(d) Random Forest has an overall accuracy of 98%, a classification precision in infected cases (1) 95% and precision in classifying deaths (2) 95% as shown in figure 7.

```
Accuracy: 0.9879486507728583
              precision    recall  f1-score

           1       0.95      0.94      0.94
           2       0.95      0.87      0.91
```

Fig. 7. Results of Random Forest

(e) Support Vector Machine has an overall accuracy of 98%, a classification precision in infected cases (1) 94% and precision in classifying deaths (2) 95% as shown in figure 8.

```
Accuracy: 0.9869007073618025
              precision    recall  f1-score

           1       0.94      0.95      0.95
           2       0.95      0.86      0.90
```

Fig. 8. Results of Support Vector Machine

From the previous results, the RF (Random Forest) algorithm has a higher score than the other algorithms and Naive Bayes has the lowest score. Therefore, this framework choose Random Forest in the next process.

Bi-Directional GR, one of the Deep Learning approaches, have used to experiment with modifying Word Embedding by choosing Covid Word Embedding and English Word Embedding, which gives accurate results as presented in table 2 and 3.

TABLE II. RESULTS OF THE TEST SET OF ENGLISH (COVID-19) WORD EMBEDDING

Polarity	Precision	Recall	F1
POSITIVE	0.77	0.78	0.78
NEGATIVE	0.78	0.77	0.77

TABLE III. RESULTS OF THE TEST ENGLISH WORD EMBEDDING

Polarity	Precision	Recall	F1
POSITIVE	0.80	0.77	0.79
NEGATIVE	0.78	0.81	0.79

Figure 9 shows the construction process of the Bi-directional GRU sentiment analysis classification model and two pre-trained word embedding was generated by Fast-Text. First word embedding is plain English text with no related to any field, and the other is word embedding that related to Covid-19.

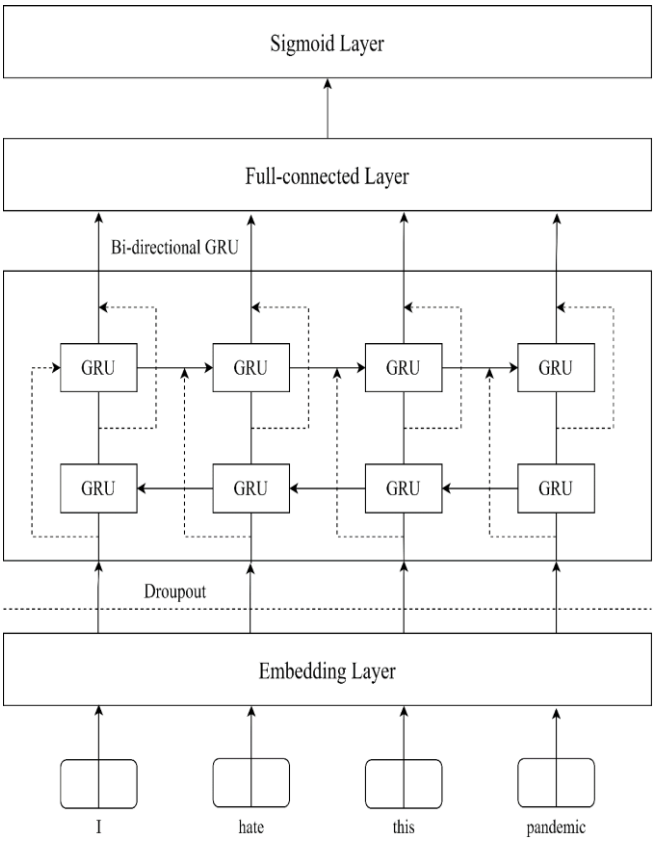

Fig. 9. Bi-directional gated recurrent neural networks (GRU) sentiment analysis model

From table 4, it shows the English word embedding has better accuracy than Covid word embedding, because our dataset (Kaggle) that we use to train is not related to Covid fields. Also, if we use the English word embedding in a real-time Tweets about Covid, it will significantly decreased the accuracy as well.

TABLE IV. RESULTS FOR INFORMATION EXTRACTION TASK

Word Embedding	Accuracy
English (Covid-19)	0.776
English	0.791

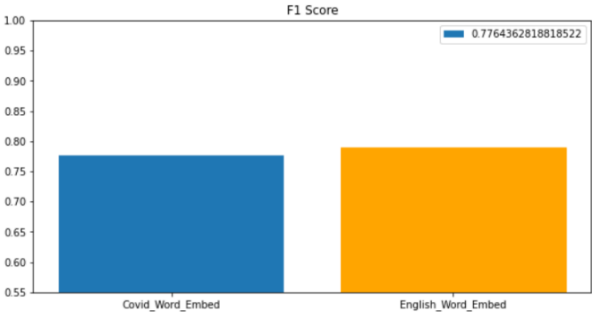

Fig. 10. Results of information Extraction

5. Conclusion

This study investigated information extraction and sentiment analysis on Twitter data. These tasks are particularly relevant when applied to social media data and the Covid19 global pandemic. The issue of information extraction on Twitter is we are labeling the data by manually unlike sentiment analysis that is Kaggle dataset. Thus, the dataset on information extraction is limited (700 tweets) not comprehensive to the other report pattern which give us limited result and accuracy. In future work, we will focus on extending and increasing the datasets of information extraction by augmentation method, and exploring more on sentiment analysis dataset in order to have more reliability in real-time use.

Acknowledgment

The authors express their sincere appreciation to Suan Sunandha Rajabhat University for financial support of the study.

References

[1] K. Chong Ng Kee Kwong, P. R. Mehta, G. Shukla, and A. R. Mehta, "COVID-19, SARS and MERS: A neurological perspective," Journal of Clinical Neuroscience, May 2020. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0967586820311851

[2] Ravi K., Ravi V., A survey of opinion mining and sentiment analysis: Tasks, approaches and applications, Knowledge-Based Systems (89) (2017), pp. 14-46

[3] Kunyanuth Kularbphettong, The awareness of environment conservation based on opinion data mining from social media, International Journal of GEOMATE, Sept., 2019 Vol.17, Issue 61, pp. 74 – 79

[4] Mihai Dusmanu, Elena Cabrio, and Serena Villata. Argument mining on twitter: Arguments, facts and sources. In EMNLP, pages 2317–2322, 2017

[5] Lara Tavoschi, Filippo Quattrone, Eleonora D'Andrea, Pietro Ducange, Marco Vabanesi, Francesco Marcelloni & Pier Luigi Lopalco (2020) Twitter as a sentinel tool to monitor public opinion on vaccination: an opinion mining analysis from September 2016 to August 2017 in Italy, Human Vaccines & Immunotherapeutics, 16:5, 1062-1069, DOI: 10.1080/21645515.2020.1714311

[6] Villavicencio, C.; Macrohon, J.J.; Inbaraj, X.A.; Jeng, J.-H.; Hsieh, J.-G. Twitter Sentiment Analysis towards COVID-19 Vaccines in the Philippines Using Naïve Bayes. Information 2021, 12, 204. https://doi.org/10.3390/info12050204

[7] Park S, Kim Y. 2016. Building thesaurus lexicon using dictionary-based approach for sentiment classification. In: 2016 IEEE 14th International Conference on Software Engineering Research, Management and Applications (SERA). Piscataway: IEEE, 39–44.

[8] Tang B, Kay S, He H. 2016. Toward optimal feature selection in naive bayes for text categorization. IEEE Transactions on Knowledge and Data Engineering 28(9):2508–2521 DOI 10.1109/TKDE.2016.2563436.

[9] M. AlRubaian, M. Al-Qurishi, M. Al-Rakhami, S. M. M. Rahman, and A. Alamri, A Multistage Credibility Analysis Model for Microblogs, presented at the Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015, Paris, France, 2015

[10] Akhtar MS, Kumar A, Ghosal D, Ekbal A, Bhattacharyya P. 2017. A multilayer perceptron based ensemble technique for fine-grained financial sentiment analysis. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. 540–546.

[11] [Adeel A, Gogate M, Hussain A. Contextual deep learning-based audio-visual switching for speech enhancement in real-world environments. Information Fusion 2020 Jul;59:163-170. [CrossRef]

[12] [Tweepy G.e.(2020),Retrieved 2021, from Tweepy: https://www.tweepy.org/

[13] DictVectorizer, Retrieved 2021, from scikit-learn.org: https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.DictVectorizer.htmlAuthor No.1, Author No 2 Onward, "Paper Title Here", Proceedings of xxx Conference or Journal (ABCD), Institution name (Country), February 21-23, year, pp. 626-632.