THAI FAKE NEWS CLASSIFICATION

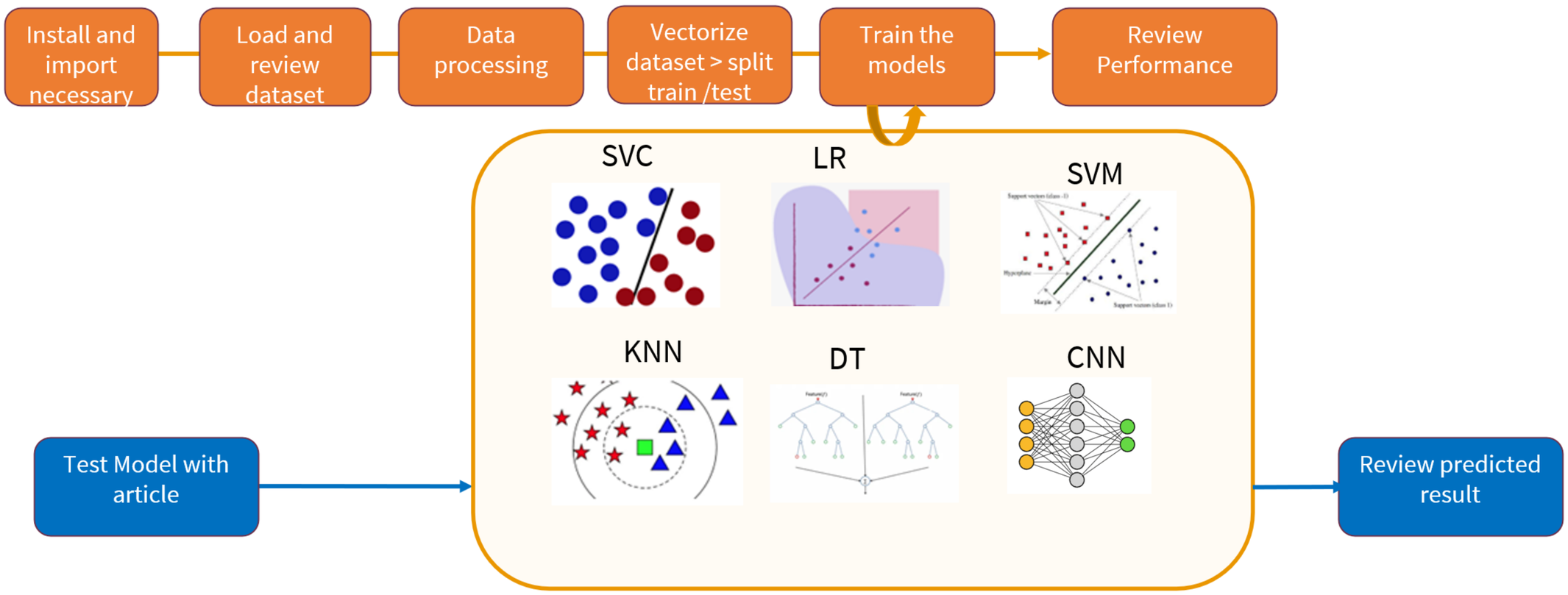
Topics List

- Project Concept and Model
- Approaches Applied
- Model Development and Demo
- Further Development

Project Concept and Model

Implement Thai Fake News Dataset in healthcare domain to classify fake news or fact news with multi-classification algorithms

Project Concept and Model



Approaches Applied

Text Cleaning

Use Regular Expression to remove special characters
Remove numbers and non-Thai characters

Removing Stop words

Loop in each documents to remove Thai stop words.
The library of the stop word downloaded from *pythainlp.corpus.common.thai_stopwords()*.

Vectorize the text using TF-IDF

Convert a collection of raw documents to a matrix of TF-IDF features.
Count Vectorizer give number of frequency with respect to index of vocabulary whereas tf-idf consider overall documents of weight of words

Text Normalization

Normalize and clean Thai text with normalizing rules, removing some texts not necessary or duplicated

Tokenize Text

Tokenizers divide strings into lists of substrings. For example, tokenizers can be used to find the words and punctuation in a string

Classification Model

- Linear SVC
- Logistic regression
- SVM
- K-Nearest Neighbors
- Decision Tree

Neural Network

- Convolutional neural networks (CNN)

*Note: See more detail in Appendix at the end of slide.

Model Development and Demo

Further Development

Idea 1 :
Apply fake news detection model to another domain

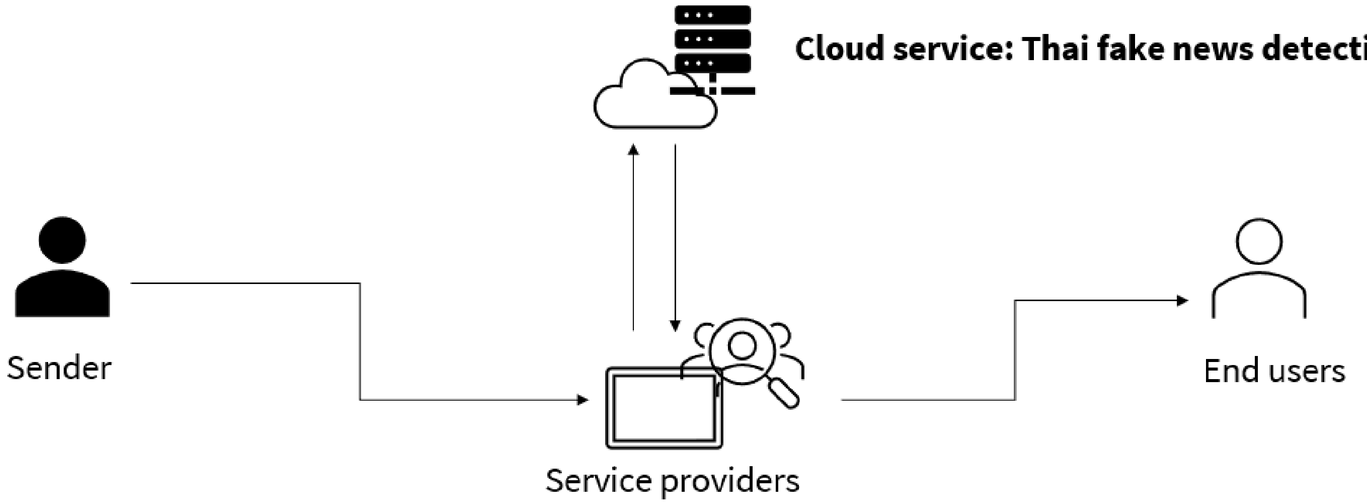
- Political news
- Money stolen message

Further Development

Idea 2 :
Thai fake new detection model (Cloud service)

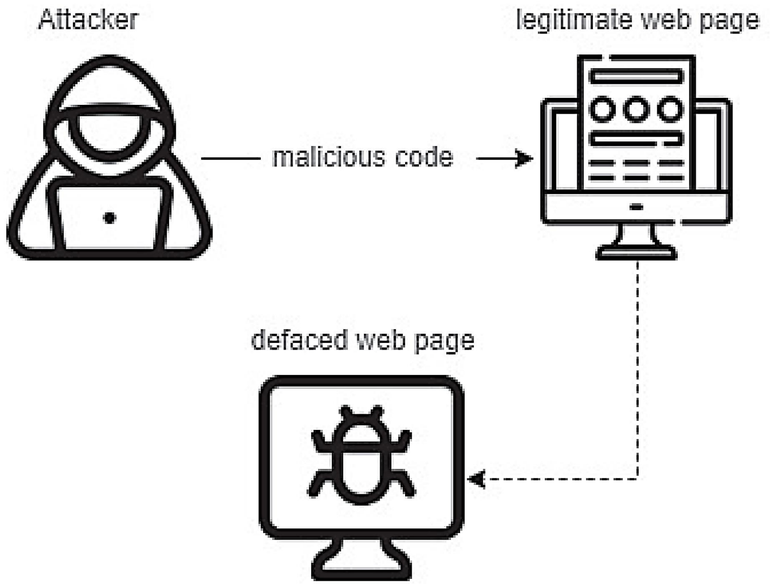
Other channels able to connect to this service to filter out fake news before sending message to end user.

- SMS Provider
- Messaging service e.g. Line



Further Development

Idea 3 :
Web Application for detecting fake news links.



Q & A

Thank You

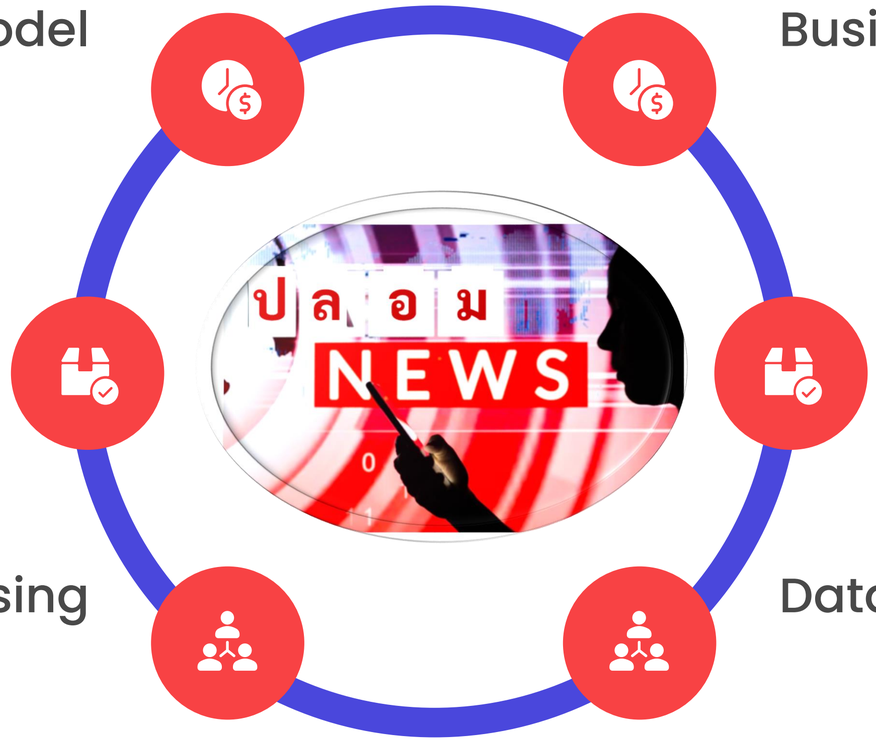
Pattareeya	Nurltes	65076051
Sasiruch	Jirasavetakul	65076064
Sekson	Hompangwhai	65076069
Apichai	Siangchin	65076074

Appendix

Methodologies

Evaluation Model

Business Understanding



Modeling

Data Understanding

Data Preprocessing

Data Preparation

1. Text cleaning

- Use Regular Expression to remove special characters
- Remove numbers and non-Thai characters

2.Text Normalization

Normalize and clean Thai text with normalizing rules as follows:

- Remove zero-width spaces
- Remove duplicate spaces
- Reorder tone marks and vowels to standard order/spelling
- Remove duplicate vowels and signs
- Remove duplicate tone marks
- Remove dangling non-base characters at the beginning of text

3.Text Tokenize

Tokenizers divide strings into lists of substrings. For example, tokenizers can be used to find the words and punctuation in a string

4.Removing Stop words

Loop in each documents to remove Thai stop words. The library of the stop word downloaded from pythainlp.corpus.common thai_stopwords().

The thai_stopwords function will return a frozen set of Thai stopwords

5.Vectorize Text Using TF-IDF

- Convert a collection of raw documents to a matrix of TF-IDF features.
- Count Vectorizer give number of frequency with respect to index of vocabulary whereas **tf-idf** consider overall documents of weight of words

smooth_idf	Smooth idf weights by adding one to document frequencies, as if an extra document was seen containing every term in the collection exactly once. Prevents zero divisions.
use_idf	Enable inverse-document-frequency reweighting. If False, idf(t) = 1.

6. Modeling

Classification Models

- Linear SVC
- Linear Regression
- Support Vector Machine (SVM)
- K-Nearest Neighbors
- Decision Tree

Neural Network Model

- Convolutional neural networks (CNN)