Misinformation Detection in Thai-Language Content on Social Media: A Case Study in Health Information

WEERANUCH PROYSAITHONG

AN INDEPENDENT STUDY SUBMITED IN PARTIAL FULFILLMENT OF THE REQUIREMENT FOR THE DEGREE OF MASTER OF SCIENCE IN DATA SCIENCE AND ANALYTICS

KMITL-DIGITAL ANALYTICS AND INTELLIGENCE CENTER SCHOOL OF SCIENCE
KING MONGKUT'S INSTITUTE OF TECHNOLOGY LADKRABANG

2024
KMITL-2024-SC-M-017-014

Independent Study Title	Misinformation Detection in Thai-Language Content on Social Media: A Case Study in Health Information
Student Name	Miss Weeranuch Proysaithong
Student ID	65056078
Degree	Master of Science (Data Science and Analytics)
	KMITL-Digital Analytics and Intelligence Center
Year	2024
Independent Study Advisor	Dr. Jiraphat Yokrattanasak

ABSTRACT

The continuous increase in the use of social media has made misinformation a significant problem. This research aims to develop a model for detecting Thai misinformation on social media, focusing on the impact of text preprocessing techniques on model accuracy. Three text preprocessing methods were selected: Replacing Numbers with Thai Words, Removing Punctuation, and Removing Stop Words. These methods were experimented with three classification models: Naïve Bayes, Support Vector Machine, and Random Forest. The results showed that the best performance was achieved when using the Remove Numbers with Thai words preprocessing method and the Support Vector Machine model, with an accuracy of 95.47%.

Keyword: Fake news detection, Text preprocessing, Thai language, Machine learning models, Social media

ACKNOWLEDGEMENT

This research was conducted to develop a model for detecting Thai misinformation on social media, focusing on the impact of text preprocessing techniques on model accuracy. The successful completion of this independent study was made possible by the guidance and advice from several instructors, to whom I express my gratitude.

I would like to thank Dr. Jiraphat Yokrattanasak for the honor of being the advisor for this independent study. Dr. Jiraphat Yokrattanasak dedicated time to provide guidance and valuable suggestions throughout the research process, from the initial stage of selecting the topic to the successful completion of the study. Dr. Jiraphat Yokrattanasak also helped identify and rectify various shortcomings, ensuring the completeness of this independent study.

I am grateful to the independent study examination committee for taking the time to listen and provide feedback, further enhancing the comprehensiveness of this research.

Finally, I would like to express my gratitude to my family and those around me who have supported and contributed to the successful completion of this independent study.

Weeranuch Proysaithong

TABLE OF CONTENTS

TABLE OF CONTENTS (Cont.)

TABLE OF CONTENTS (Cont.)

LIST OF TABLES

Page

LIST OF FIGURES

Page

Chapter 1
Introduction

Introduction will provide an overview of the study, including the background and significance of the research, the objectives it aims to achieve, the scope and limitations of the study, and the anticipated benefits upon successful completion. This chapter will establish the context and relevance of the research, guide the study's direction, acknowledge any constraints, and justify the importance of the findings and their potential contributions to the field.

1.1 Statement and Significance of The Problems

According to the 2023 Anti-Fake News Center's performance report, the health category had the highest number of fake news stories, with 2,003 articles. Statistical data trends suggest that the health category will involve various issues, such as fake news about emerging or severe epidemic diseases, exaggerated claims about health products, news about health, diseases, and herbal treatments, and health news related to hot weather and heat stroke.

Table 1 News Dissemination Results Categorized by News Category

News Category	Factual News	Fake News	Misleading News
Health Category	153 articles	2,003 articles	210 articles
Disaster Category	47 articles	204 articles	32 articles
Government Policy Category	1,101 articles	1,513 articles	206 articles
Economy Category	46 articles	835 articles	40 articles

Remark: Data from November 1, 2019 to August 31, 2023 [1]

As people from various social groups increasingly consume health information via web-based platforms, their exposure to health misinformation rises. Although these platforms can be valuable for health promotion, they can spread false and misleading health information faster than scientific knowledge, raising serious public health concerns. [2]

This research will contribute to the development of an effective model for detecting Thai health-related misinformation on social media. Furthermore, text preprocessing is a crucial step in developing an efficient model. This study will help us understand the impact of text preprocessing techniques on the development of misinformation detection models. The findings from this research can be used as a decision-support tool for stakeholders, including general social media users and researchers interested in developing misinformation detection models. Ultimately, this research will expand knowledge in the field of misinformation detection, focusing on the impact of text preprocessing, which will promote future development and research.

1.2 Goal and Objective

The primary goal of this research is to investigate the impact of three text preprocessing methods on the performance of machine learning models for text classification. The specific objectives are:

1. To evaluate the effect of Replacing Numbers with Thai Words, Removing Punctuation, and Removing Stop Words on the accuracy of Naïve Bayes, Support Vector Machine, and Random Forest models.
2. To determine which individual preprocessing method or combination of methods yields the highest accuracy for each machine learning model.

By comparing the results of different preprocessing techniques, this research aims to identify the most effective approach for enhancing the performance of text classification models in Thai health-related news articles from social media.

1.3 Scope of The Study

1. The data used in this research will be Thai health-related news articles from social media.
2. This study will focus on investigating the impact of text preprocessing techniques relevant to text classification tasks.

1.4 Expected Benefits

1. **Understanding the impact of text preprocessing:** The findings of this research will help us understand the impact of text preprocessing techniques on the development of misinformation detection models, which will contribute to the future development and improvement of text preprocessing techniques.
2. **Prevention and management of misinformation:** The results of this study will provide us with an effective tool for detecting and managing Thai health-related misinformation on social media, which will help reduce the impact caused by misinformation.
3. **Enhancing Python programming skills:** Through this research process, you will gain experience in using Python for research purposes, including text preprocessing and the development of machine learning model.

Chapter 2
Literature Review

This chapter will cover relevant theories and research related to the topic of this study. It will provide an overview of the existing knowledge and identify the gaps in the current literature.

2.1 Conceptual Framework

The relevant theories in this section will be divided into two parts. The first part will focus on theories related to the definitions and types of misinformation. The second part will cover theories associated with machine learning, specifically those pertaining to classification tasks. This section will also include a discussion on evaluating models using confusion matrices. By presenting these theories, the conceptual framework will provide a clear understanding of the key concepts and methods that form the foundation of this research.

2.1.1 Definitions of Misinformation, Disinformation, and Fake News

In the realm of media and communication, the terms misinformation, disinformation, and fake news [3-7] are often used interchangeably, but they have distinct meanings. **Misinformation** refers to false information that is spread, regardless of whether there is an intent to mislead. On the other hand, **Disinformation** is deliberately misleading or biased information, manipulated narratives, or facts, often in the form of propaganda. **Fake news**, a term that has gained prominence in recent years, is defined as false stories that appear to be news, spread on the internet or using other media, usually created to influence political views or as a joke. It is also described as purposefully crafted, sensational, emotionally charged, misleading, or totally fabricated information that mimics the form of mainstream news

Understanding these distinctions is crucial in combating the spread of false information and promoting media literacy in today's digital age.

2.1.2 Machine Learning for Classification Tasks

Machine Learning (ML), a subset of Artificial Intelligence, originated from pattern recognition and the theory that computers can learn and gradually improve their performance without being explicitly programmed. This concept sparked the interest of AI researchers in exploring whether computers and systems can learn from interacting with data. A key characteristic of ML is its iterative process, which is crucial for developing capabilities. The systems and analytical models adapt to the datasets they encounter, leading to self-improvement. The system learns and corrects errors from previous tasks until it can produce reliable and consistent results, demonstrating the ability to learn from pre-prepared training data.

Machine learning can be divided into several main groups based on the type of training: Supervised Learning, Unsupervised Learning, Semi-supervised Learning, and Reinforcement Learning. Supervised learning can be further categorized into two types: Regression and Classification. Creating misinformation detection models falls under the classification category, which includes various techniques such as Naïve Bayes, Support Vector Machines, and Random Forests. These techniques will be discussed in more detail in the following sections.

2.1.2.1 Naïve Bayes

Naive Bayes [8-10] classifiers are a family of probabilistic machine learning models widely used for classification tasks such as text classification, spam filtering, and recommendation systems. These classifiers are based on Bayes' theorem, which calculates the probability of a given input belonging to a particular class, assuming independence between features. This simplifying assumption, although often not true in real-world scenarios, allows Naive Bayes classifiers to make quick and accurate predictions.

As part of the generative learning algorithms family, Naive Bayes classifiers model the distribution of inputs for a given class or category, rather than learning which features are most important for differentiation between classes. There are several types of Naive Bayes classifiers, each suited for different data distributions and use cases. Gaussian Naive Bayes (GaussianNB) is used with continuous variables and assumes a normal distribution, while Multinomial Naive Bayes (MultinomialNB) is applied to discrete data, such as frequency counts in natural language processing. Bernoulli Naive Bayes (BernoulliNB) is used with Boolean variables, making it suitable for binary classification tasks.

Despite their simplicity, Naive Bayes classifiers offer several advantages. They are less complex compared to other classifiers, making them easier to understand and implement. These classifiers also scale well, providing fast and efficient performance when the conditional independence assumption holds. Additionally, Naive Bayes classifiers can handle high-dimensional data, such as in document classification tasks, where other classifiers may struggle.

However, Naive Bayes classifiers also have some limitations. They are subject to the "Zero Frequency" problem, which occurs when a categorical variable is not present in the training data, leading to a zero probability and potentially incorrect classifications. This issue can be mitigated using smoothing techniques like Laplace estimation. Furthermore, the core assumption of feature independence is often unrealistic in real-life scenarios, which can lead to suboptimal performance.

Bayes theorem provides a way of computing posterior probability $P(c \mid x)$ from $P(c)$, $P(x)$ and $P(x \mid c)$. Look at the equation below [8]:

$$P(c \mid x) = \frac{P(x \mid c)P(c)}{P(x)}$$

$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid \times P(c) \qquad (1)$$

Above,

$P(c \mid x)$ is the posterior probability of class (c, target) given predictor (x, attributes).

$P(c)$ is the prior probability of class.

$P(x \mid c)$ is the likelihood which is the probability of the predictor given class.

$P(x)$ is the prior probability of the predictor.

In conclusion, Naive Bayes classifiers are powerful tools in machine learning, offering a probabilistic approach to classification tasks. Despite their simplifying assumptions, these classifiers have proven to be effective in various domains, including text classification, spam filtering, and sentiment analysis. By understanding the strengths and limitations of Naive Bayes classifiers, researchers and practitioners can effectively apply these models to solve real-world problems.

2.1.2.2 Support Vector Machine

Support Vector Machine (SVM) [11-13] is a supervised machine learning algorithm that has gained significant popularity due to its ability to efficiently classify data points in high-dimensional spaces. Developed by Vladimir N. Vapnik and his colleagues in the 1990s, SVM has become a go-to algorithm for various classification problems, particularly in text classification tasks.

The primary objective of SVM is to find an optimal hyperplane that maximizes the margin between different classes of data points in an N-dimensional space. In an SVM model, each data point is represented as a point in n-dimensional space, where n is the number of features. The algorithm then seeks to find the hyperplane that provides the maximum margin of separation between the two classes. This hyperplane is defined by the support vectors, which are the data points closest to the decision boundary, as shown in Figure 1 [11].

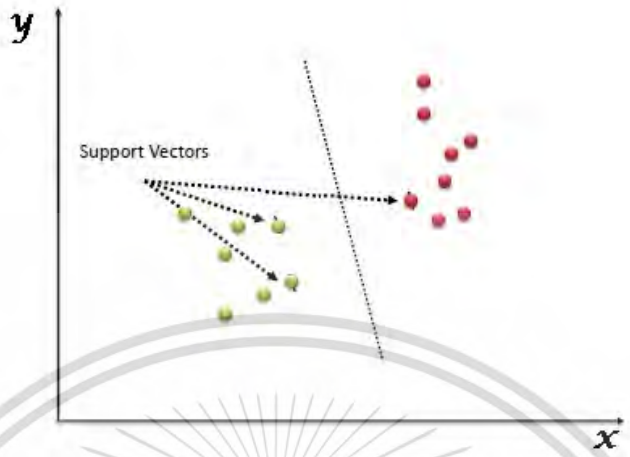

Figure 1 Support Vectors Defining the Decision Boundary

The hyperplane's dimension is determined by the number of input features, with a line representing the hyperplane in a 2-D space and a plane in higher-dimensional spaces. By maximizing the margin, SVM ensures that the decision boundary is as far away as possible from the closest data points of each class, known as support vectors. This approach enables SVM to generalize well to new data and make accurate predictions. One of the key strengths of SVM is its ability to handle both linear and nonlinear classification tasks. When the data is not linearly separable, kernel functions, such as linear, polynomial, radial basis function (RBF), or sigmoid kernels, can be employed to transform the data into a higher-dimensional space where linear separation becomes possible. This technique, known as the "kernel trick," allows SVM to tackle complex classification problems effectively.

However, SVM also has its limitations. It may not perform well when dealing with large datasets, as the training time can be significantly higher. Additionally, SVM's performance can be impacted by the presence of noise or overlapping target classes in the dataset. Another drawback is that SVM does not directly provide probability estimates, requiring an expensive five-fold cross-validation process to calculate them.

Despite these limitations, SVM remains a powerful and widely used algorithm, particularly for building machine learning models with small datasets. Its effectiveness in high-dimensional spaces and its ability to handle cases where the number of

dimensions exceeds the number of samples make it a valuable tool in various domains.

To build an SVM classifier, the first step is to split the data into training and testing sets. Hyperparameters can be tuned using techniques like grid search and cross-validation to improve the model's performance. When compared to other supervised learning classifiers, such as Naive Bayes, SVM often outperforms in scenarios where the data is not linearly separable, although it may be more computationally expensive and require more hyperparameter tuning.

In conclusion, Support Vector Machine is a robust and versatile supervised learning algorithm that excels in classification tasks, particularly when dealing with high-dimensional data and small datasets. Its ability to find optimal decision boundaries and generalize well to new data makes it a valuable tool in various industries and applications.

2.1.2.3 Random Forest

Random Forest [14-16] is a powerful and widely-used machine learning algorithm that combines multiple decision trees to create a more accurate and robust predictive model. Developed by Leo Breiman and Adele Cutler, Random Forest is known for its ease of use and flexibility, as it can handle both classification and regression problems effectively.

At its core, Random Forest is an ensemble learning method that leverages the power of multiple decision trees. Decision trees are simple yet effective algorithms that make predictions based on a series of questions or decision nodes. These nodes split the data based on various criteria, such as the presence of a specific feature or the value of a variable. The final prediction is made at the leaf node, which represents the outcome of the decision-making process.

While decision trees are useful, they can be prone to overfitting and bias. This is where Random Forest comes in. By creating an ensemble of decision trees and aggregating their predictions, Random Forest mitigates these issues and produces more

accurate results. The algorithm introduces randomness in two ways: by using a random subset of the training data for each tree (bagging) and by considering a random subset of features at each decision node (feature bagging). The Random Forest algorithm works as follows:

1. A random subset of the training data is selected with replacement (bootstrap sample) for each decision tree.
2. At each decision node, a random subset of features is considered for splitting the data.
3. The decision trees are grown independently, without pruning, to their maximum depth.
4. For classification tasks, the final prediction is determined by majority voting among the trees. For regression tasks, the average of the individual tree predictions is used.

One of the key advantages of Random Forest is its ability to handle large and complex datasets with both continuous and categorical variables. It is also less susceptible to overfitting compared to individual decision trees, thanks to the randomness introduced in the training process. Additionally, Random Forest provides a measure of feature importance, allowing users to identify the most influential variables in the model.

However, Random Forest does have some limitations. The algorithm can be computationally expensive, especially when dealing with a large number of trees or features. It may also require more memory and storage resources compared to simpler models. Furthermore, the interpretability of Random Forest models can be challenging, as the final prediction is based on the collective output of multiple decision trees.

Despite these challenges, Random Forest remains a popular choice among data scientists and machine learning practitioners. Its versatility, robustness, and relatively low hyperparameter tuning requirements make it a go-to algorithm for various applications, including image classification, fraud detection, and customer churn prediction.

In conclusion, Random Forest is a powerful and flexible machine learning algorithm that combines the strengths of multiple decision trees to create accurate and robust predictive models. By leveraging the power of ensemble learning and introducing randomness in the training process, Random Forest overcomes the limitations of individual decision trees and provides a reliable tool for both classification and regression tasks.

2.1.2.4 Confusion Matrix

A confusion matrix [17-19] is a crucial tool for evaluating the performance of a machine learning classification model. It provides a comprehensive view of the model's predictions by comparing them with the actual target values. The matrix displays the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), allowing for a detailed analysis of the model's accuracy and the types of errors it makes. An example of a 2x2 confusion matrix is shown in Table 2.

Table 2 Confusion Matrix

Class	Actual Value / Positive	Actual Value / Negative
Predicted Value / Positive	True Positives (TP)	False Positives (FP)
Predicted Value / Negative	False Negatives (FN)	True Negatives (TN)

In a confusion matrix, the columns represent the actual values of the target variable, while the rows represent the predicted values. True positives occur when the model correctly predicts a positive value, and true negatives occur when the model correctly predicts a negative value. False positives, also known as type I errors, happen when the model incorrectly predicts a positive value when the actual value is negative. False negatives, or type II errors, occur when the model incorrectly predicts a negative value when the actual value is positive.

The confusion matrix provides valuable insights beyond basic accuracy metrics, especially when dealing with imbalanced datasets. It enables the calculation of various performance metrics, such as precision, recall, F1-score, and specificity. Precision measures the accuracy of positive predictions, while recall evaluates the model's ability to identify all relevant instances. The F1-score is the harmonic mean of precision and recall, providing an overall assessment of the model's performance. Specificity, on the other hand, measures the model's ability to correctly identify negative instances.

Interpreting a confusion matrix helps in understanding the strengths and weaknesses of a classification model. It allows for the identification of areas where the model excels and where it needs improvement. By analyzing the distribution of TP, TN, FP, and FN, data practitioners can fine-tune their models to achieve better results.

The confusion matrix is particularly useful when dealing with imbalanced data, where one class significantly outnumbers the other. In such cases, relying solely on accuracy can be misleading, as the model may achieve high accuracy by simply predicting the majority class. The confusion matrix, along with metrics like precision and recall, provides a more balanced view and helps in making informed decisions.

Furthermore, the confusion matrix helps in understanding the trade-offs between different metrics. For example, increasing precision may lead to a decrease in recall, and vice versa. By examining these trade-offs, data practitioners can choose the most appropriate metric based on the specific requirements of their application.

In conclusion, the confusion matrix is an essential tool for evaluating and improving the performance of machine learning classification models. It provides a comprehensive view of the model's predictions, enables the calculation of various performance metrics, and helps in identifying areas for improvement. By leveraging the insights provided by the confusion matrix, data practitioners can develop more accurate and reliable models for a wide range of applications.

2.1.3 Text Preprocessing Techniques

In the model development process for natural language processing (NLP) and machine learning tasks involving textual data, text preprocessing is a crucial step. [20-24] The process typically begins with data collection, followed by text cleaning and then text preprocessing. After these initial steps, the focus shifts to feature engineering and model development. Finally, the results of the developed model are tested to evaluate its performance and effectiveness.

The primary goal of text preprocessing is to transform raw, unstructured text into a clean, structured format that can be effectively used for analysis and modeling. This process involves various techniques aimed at improving the quality and usability of the text data.

The main text preprocessing techniques include lowercasing, tokenization, removing punctuation and special characters, removing stop words, removing URLs and HTML tags, stemming, and lemmatization. Lowercasing involves converting all letters in the text to lowercase to ensure consistency and avoid treating the same words differently based on their case. Tokenization is the process of breaking down large blocks of text into smaller units, such as sentences or words, to facilitate analysis.

Removing punctuation, special characters, and digits is another important step in text preprocessing. These elements often do not provide meaningful information for text analysis and can interfere with NLP algorithms. Similarly, removing URLs and HTML tags is necessary when working with text data obtained from web pages or other HTML-formatted sources.

Removing stop words is a technique used to eliminate commonly occurring words that do not contribute significantly to the meaning of a sentence, such as "the," "a," "an," and "in." By removing these words, the focus can be placed on the more important and informative words in the text.

Stemming and lemmatization are two methods used to reduce words to their base or dictionary form. Stemming is a rule-based approach that removes word affixes, resulting in the word's stem. However, stemming may sometimes produce non-dictionary words. Lemmatization, on the other hand, uses a vocabulary and

morphological analysis to reduce words to their base form, ensuring that the resulting words are part of the language vocabulary. The choice between stemming and lemmatization depends on the specific requirements of the project, considering factors such as simplicity, speed, and accuracy.

In conclusion, text preprocessing is an essential step in NLP and machine learning projects involving textual data. By applying techniques such as lowercasing, tokenization, removing unwanted characters and words, stemming, and lemmatization, the quality and usability of the text data can be significantly improved. These preprocessing steps help in reducing noise, converting the data into a structured format, and focusing on the most important information, ultimately leading to better performance and accuracy of the analysis or modeling tasks.

2.2 Review of Previous Research Works

The review of related research will be divided into two parts: research on developing Thai misinformation detection models and research on the impact of text preprocessing on text classification.

2.2.1 Related Research on Developing Thai Misinformation Detection Models

Studies related to developing Thai misinformation detection models mainly focus on analyzing fake information using machine learning and deep learning techniques. The results from these studies can be summarized in Table 3.

Table 3 Related Research on Developing Thai Misinformation Detection Models

Study (Year)	Preprocessing Methods	ML Methods	Best Results
Detecting Fake News with Machine Learning Method (2018) [25]	- Text preprocessing - Normalization - Removing duplicate data	- Naïve Bayes - Neural Network - Support Vector Machine	- Support Vector Machine and Neural Network had highest accuracy for fake news detection.
The COVID-19 Fake News Detection in Thai Social Texts (2021) [26]	- Spelling corrections - Removing redundant characters - Removing punctuation - Removing punctuation	- BERT - ULMFiT - GPT	- ULMFiT showed higher performance for Thai COVID-19 fake news detection
Artificial Intelligent Techniques for Thai Fake News Detection (2022) [27]	- Word segmentation - Removing unwanted words - Keyword Extraction	- Logistic Regression - K-Nearest Neighbor - Naïve Bayes - Multilayer Perceptron - Support Vector Machine - Random Forest - LSTM	- LSTM had highest accuracy for fake news detection.

Table 3 Related Research on Developing Thai Misinformation Detection Models

Study (Year)	Preprocessing Methods	ML Methods	Best Results
Fake News Detection on Social Media: Case Study of Coronavirus 2019 (2022) [28]	- Removing duplicate and missing data - Text preprocessing - Normalization	- Decision Tree - BiLSTM - BiGRU	- Decision Tree had highest accuracy for fake news detection.
Thai Fake News Detection Using Machine Learning Model (2022) [29]	- Text preprocessing - Normalization	- Naïve Bayes - Neural Network - Support Vector Machine	- Support Vector Machine had highest accuracy for fake news detection.

2.2.2 Related Research on the Impact of Text Preprocessing on Text Classification

Research on the impact of text preprocessing on text classification primarily investigates the effects of various preprocessing methods on model accuracy using different datasets. The findings from these studies can be summarized in Table 4.

Table 4 Related Research on the Impact of Text Preprocessing on Text Classification

Study (Year)	Preprocessing Methods	ML Methods	Best Results
The Effect of Preprocessing Techniques on Twitter Sentiment Analysis (2016) [30]	- Weighting scheme - Stemming - Removing Stop Words - Tokenization - Feature selection	- Naïve Bayes - Support Vector Machine - k-Nearest Neighbor - C4.5 Decision Tree	- Unigrams and 1-3-grams best for accuracy. - Feature selection improves accuracy.
The Influence of Preprocessing on Text Classification (2020) [31]	- Spelling correction - HTML tag removal - Converting uppercase to lowercase - Removing Punctuation - Reducing repeated characters - Removing Stop Words	- Bayes Networks - Random Forest - SMO (a variant of SVM)	- Stop-word removal topped solo preprocessing methods - Spelling correction played a role in best combinations

Table 4 Related Research on the Impact of Text Preprocessing on Text Classification

Study (Year)	Preprocessing Methods	ML Methods	Best Results
Exploring the Relationship Between Algorithm Performance, Vocabulary, and Run-Time in Text Classification (2021) [32]	- Lowercasing - Rare word filtering - Hashing - Removing Punctuation - Removing Stop Words - Removing Number - Word stemming - Lemmatization - Spelling correction - Word segmentation	- k-Nearest Neighbor - Naïve Bayes - Support Vector Machine	- Stop-word removal and rare word filtering improve processing speed and accuracy.

2.3 Identification of Gaps

After reviewing the relevant research, it was found that most studies on Thai misinformation detection focus on developing models to achieve good results. However, there is a lack of research exploring the impact of text preprocessing techniques on model accuracy, particularly for the Thai language. Furthermore, studies on the impact of text preprocessing on text classification are limited when it comes to Thai data. Therefore, this research aims to develop a model for detecting Thai misinformation on social media, with a focus on the impact of text preprocessing techniques on model accuracy.

Chapter 3
Research Methodology

The methodology section will begin by providing an overview of the entire process and then delve into the details of each step. This research aims to develop a model for detecting Thai misinformation on social media, emphasizing the impact of text preprocessing techniques on model accuracy. The overall research steps are illustrated in Figure 2.

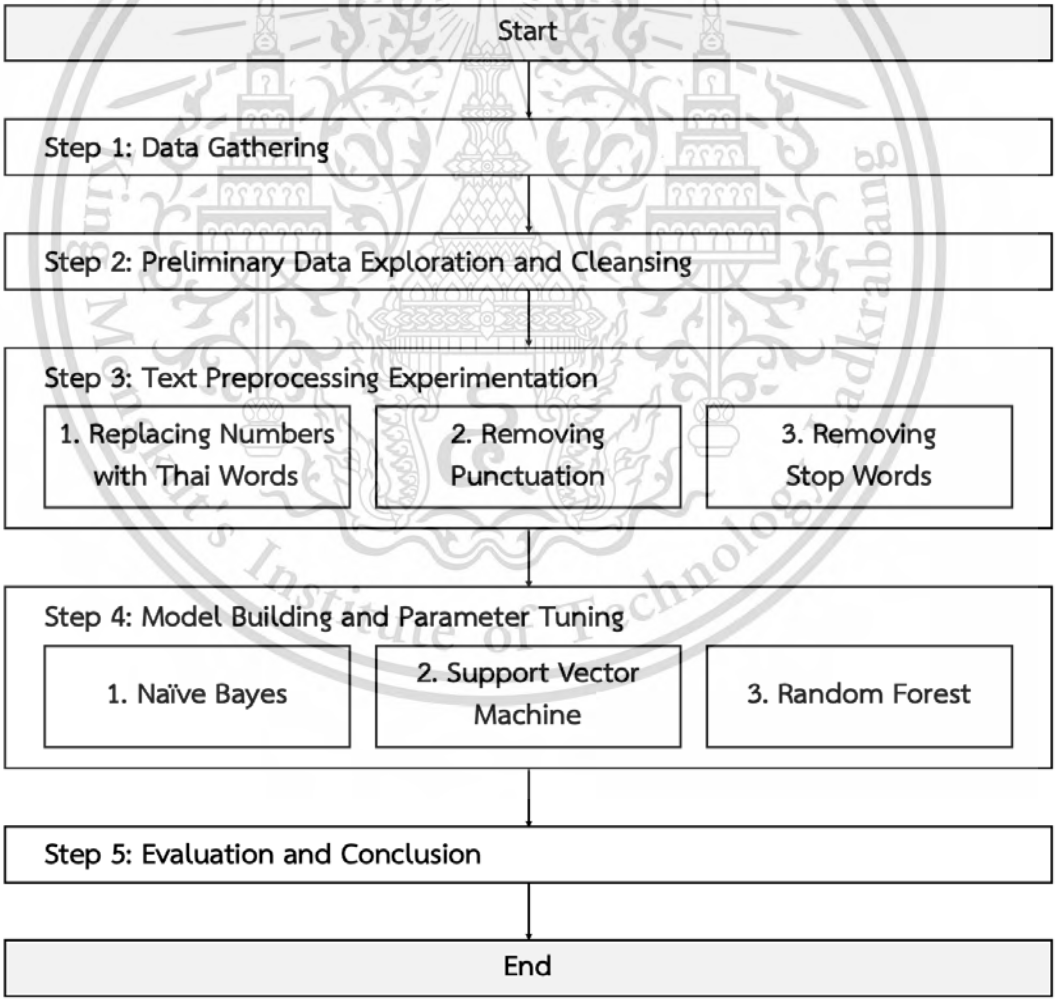

Figure 2 Research Methodology

3.1 Data Gathering

Data for this research was collected from two main reliable sources for Thai health-related information: fake data and factual data.

Initially, representative fake and factual data were gathered from the open data provided by the Anti-Fake News Center Thailand, accessed through the website https://opendata.antifakenewscenter.com. Data from the health products category, spanning from March 2017 to February 2024, was selected. There are examples of real and fake data as shown in Figure 3.

Figure 3 Examples of real and fake news

However, the proportion of factual data obtained was insufficient for use. Subsequently, additional factual health data was collected from the Thai Health Promotion Foundation (ThaiHealth) via the website https://www.thaihealth.or.th. Data from the health news category, ranging from November 2018 to February 2024, was chosen. After obtaining data from both sources, it was combined into a single dataset. The details of the column names and their descriptions are presented in Table 5.

Table 5 Column Names and Details in the Dataset

No.	Column	Column Details
1	topic	News headline
2	label	Classification of news (factual/fake)

3.2 Preliminary Data Exploration and Cleansing

Following the data collection from the two sources mentioned in section 3.1, initial data cleaning was performed by removing duplicate data, checking for null values, and removing irrelevant non-health-related data. Additionally, words indicating fake news in the news headlines were removed. For example, the headline "ข่าวปลอม อย่าแชร์! เครื่องจัดฟันซิลิโคน สามารถใช้จัดฟันเองได้ โดยไม่ต้องไปพบแพทย์" was changed to "เครื่องจัดฟันซิลิโคน สามารถใช้จัดฟันเองได้ โดยไม่ต้องไปพบแพทย์"

An exploration of the data revealed a significant disparity in the number of records between the factual and fake news classes. To prevent the issue of imbalanced data, random sampling was performed to reduce the data to 1,600 records per class, as shown in Table 6.

Table 6 Number of Records Before and After Handling Imbalanced Data

News Category	Number of Records Before Handling Imbalanced Data	Number of Records After Handling Imbalanced Data
Fake News	1,696	1,600
Factual News	4,246	1,600

3.3 Text Preprocessing Experimentation

After initial data cleaning, further preprocessing is necessary to improve model performance. The preprocessing methods are divided into essential techniques, including tokenization and vectorization. For tokenization, the word_tokenize library from PyThaiNLP with the engine="newmm" was used, and vectorization was performed using the TF-IDF method.

PyThaiNLP is a popular library for Thai natural language processing tasks using Python. It provides functions such as word tokenization, part-of-speech tagging, transliteration, soundex generation, and spell checking.

Additionally, three text preprocessing methods were selected for experimentation: Replacing Numbers with Thai Words, Removing Punctuation, and Removing Stop Words. The aim is to determine which method or combination of methods yields the highest model accuracy, as increasing the number of preprocessing steps may not always improve model performance. The details of each method are as follows:

3.3.1 Replacing Numbers with Thai Words

Upon examining the data, it was found that the news headlines contained numeric information. In this step, the data is preprocessed by converting numbers to Thai words using the num2words library from PyThaiNLP. After this step, numbers are converted to words. For example, '19' is transformed into 'สิบเก้า'.

3.3.2 Removing Punctuation

Data exploration revealed the presence of punctuation in the news headlines. This step involves preprocessing the data by removing specified Thai punctuation marks from the text. The Thai punctuation marks used are based on “วารสารราชบัณฑิตยสถาน”, as shown in Table 7.

Table 7 Thai Punctuation Marks

No.	Thai Name	English Name	Symbol
1	มหัพภาค	full stop, period	.
2	จุลภาค	comma	,
3	อัฒภาค	semicolon	;
4	ทวิภาค	colon	:
5	วิภัชภาค	colon and dash	:-
6	ยัติภังค์	hyphen	-

Table 7 Thai Punctuation Marks

No.	Thai Name	English Name	Symbol
7	วงเล็บ หรือ นขลิขิต	parentheses	( )
8	วงเล็บเหลี่ยม	square brackets	[ ]
9	วงเล็บปีกกา	braces	{ }
10	ปรัศนี	question mark	?
11	อัศเจรีย์	exclamation mark	!
12	อัญประกาศ	double quotation marks	“ ”
13	อัญประกาศเดี่ยว	single quotation marks	‘ ’
14	ไม้ยมก หรือ ยมก	-	ๆ
15	ปยาลน้อย หรือ เปยยาลน้อย	-	ฯ
16	ไปยาลใหญ่ หรือ เปยยาลใหญ่	et cetera, etc.	ฯลฯ
17	ไข่ปลา หรือ จุดไข่ปลา	ellipsis, dotted line	.....
18	เส้นประ	dash line	– – –
19	เสมอภาค หรือ เท่ากับ	equals	=
20	สัญประกาศ	underline	___
21	บุพสัญญา	ditto mark	”
22	มหัตถสัญญา	paragraph	¶
23	ดอกจัน	remark, note	*
24	ทับ	virgule, slant, slash	/

3.3.3 Removing Stop Words

This step preprocesses the data by removing stop-words, which are common words frequently found in sentences but do not convey significant meaning, such as "การ", "ความ", "คือ", "ที่", and "ซึ่ง". The tokenization process is also performed in this step. The stop-words used are from PyThaiNLP, and examples of Thai stop-words are illustrated in Figure 4.

```
List of Thai stop words (concise):
สมัยโน้น
พร้อมทั้ง
เกิด
ภายภาค
ยังแต่
เสมือนกับ
มี
เช่นเคย
จะได้
นอกจากนั้น
นิดหน่อย
เพียงแค่
พอแล้ว
เห็นว่า
เมื่อนั้น
เหตุนั้น
แต่ถ้า
พร้อมกัน
ครั้ง
ปิด
ตาม
น่ะ
ทุกตัว
ที่จริง
ความ
คราวไหน
แห่งใด
ทุกสิ่ง
อดีต
ไง
อย่างไรก็ได้
ยิ่งใหญ่
เกินๆ
อาจเป็นด้วย
แต่นั้น
```

Figure 4 Examples of Thai Stop Words from thai_stopwords

3.3.4 All Cases for Experimenting with Changes in Text Preprocessing Methods

The experimentation process for changing text preprocessing methods involves testing a total of 8 cases, ranging from not using any of the three methods to using all three methods. The cases are summarized in Table 8.

Table 8 All Cases for Experimenting with Changes in Text Preprocessing Methods

Case	Method	Replacing Numbers with Thai Words	Removing Punctuation	Removing Stop Words
1	No	No	No	No
2	1	Yes	No	No
3	2	No	Yes	No
4	3	No	No	Yes
5	1,2	Yes	Yes	No
6	1,3	Yes	No	Yes
7	2,3	No	Yes	Yes
8	1,2,3	Yes	Yes	Yes

Remark: 1 = Replacing Numbers with Thai Words, 2 = Removing punctuation, 3 = Removing stop-words

3.4 Model Building and Parameter Tuning

In the experiment of building models for classifying misinformation, three classification models were selected for development: Naïve Bayes, Support Vector Machine, and Random Forest. The aim is to compare how different text preprocessing methods affect the accuracy of each model. The experiment specifies the architecture and parameter tuning for each model, as indicated in Table 9.

Table 9 Tuning Parameters for Each Model

Model	Parameter Tuning
Naïve Bayes	Type of Naïve Bayes model = Multinomial Naïve Bayes
Support Vector Machine	Kernel = linear C = 1 gamma = 0
Random Forest	n_estimators=200, max_depth=20, min_samples_leaf=5

3.5 Evaluation and Conclusion

The accuracy of different cases is compared using the accuracy of each model. Further analysis is conducted to determine the reasons behind the superior performance of certain models in specific cases.

Chapter 4
Results

In this research, the results are divided into two parts: the impact of different text preprocessing methods on data cleaning and the accuracy of each model when using various text preprocessing techniques.

4.1 Text Preprocessing Results

The text preprocessing experiments involve eight cases, ranging from not using any of the three preprocessing methods (Replacing Numbers with Thai Words, Removing Punctuation, Removing Stop Words) to using all three methods, as specified in section 3.4.4 The results of the data cleaning are as follows:

4.1.1 Text Preprocessing Results for Case 1

This case involves cleaning the data without using any of the three text preprocessing methods (Replacing Numbers with Thai Words, Removing Punctuation, Removing Stop Words). Examples of the data before and after cleaning, as specified, are shown in Table 10.

Table 10 Text Preprocessing Results for Case 1

No.	Topic	Topic_After	Label
0	กรมควบคุมโรค กระทรวงสาธารณสุข ลงพื้นที่สนามบินดอนเมืองตรวจเยี่ยมประชาสัมพันธ์ และบังคับใช้กฎหมายสถานที่สาธารณะปลอดบุหรี่ พบนักท่องเที่ยว และประชาชนผู้ใช้บริการให้การตอบรับและยินยอมปฏิบัติตามกฎหมายสถานที่สาธารณะปลอดบุหรี่ฉบับใหม่ ซึ่งมีผลตั้งแต่วันที่ 3 กุมภาพันธ์ 2562 ที่ผ่านมา	กรม ควบคุม โรค กระทรวง สาธารณสุข ลง พื้นที่ สนามบิน ดอน เมือง ตรวจ เยี่ยม ประชาสัมพันธ์ และ บังคับใช้ กฎหมาย สถานที่สาธารณะ ปลอด บุหรี่ พบ นักท่องเที่ยว และ ประชาชน ผู้ใช้บริการ ให้ การตอบรับ และ ยินยอม ปฏิบัติ ตามกฎหมาย สถานที่สาธารณะ ปลอด บุหรี่ ฉบับ ใหม่ ซึ่ง มีผล ตั้งแต่ วันที่ 3 กุมภาพันธ์ 2562 ที่ผ่านมา	ข่าวจริง
1	ยาป้องกันและรักษาโรคไวรัสปอดอักเสบ COVID-19	ยา ป้องกัน และ รักษาโรค ไวรัส ปอด อักเสบ COVID- 19	ข่าวปลอม
2	สมุนไพรพิลังกาสา รักษาโรคมะเร็ง	สมุนไพร พิลังกาสา รักษา โรคมะเร็ง	ข่าวปลอม
3	ไข้เลือดออก ระบาด "ชัยภูมิ-ระยอง-ขอนแก่นแม่ฮ่องสอน-นครราชสีมา" อัตราป่วยสูงสุด ทั่วไทยพบป่วย 2.5 หมื่นคน เสียชีวิต 15 ราย ปัจจุบันเพิ่มความเสี่ยง หลัง อุตุฯพยากรณ์ ฝนจะตกหลายพื้นที่ งัดมาตรการ 3 เก็บออกมารับมือ	ไข้เลือดออก ระบาด " ชัยภูมิ - ระยอง - ขอนแก่น แม่ฮ่องสอน - นครราชสีมา " อัตรา ป่วย สูงสุด ทั่ว ไทย พบ ป่วย 2.5 หมื่น คน เสียชีวิต 15 ราย ปัจจุบัน เพิ่ม ความเสี่ยง หลัง อุตุฯ พยากรณ์ ฝน จะ ตก หลาย พื้นที่ งัด มาตรการ 3 เก็บ ออกมา รับมือ	ข่าวจริง

Table 10 Text Preprocessing Results for Case 1

No.	Topic	Topic_After	Label
4	กรมวิทยาศาสตร์การแพทย์ประกาศความสำเร็จการทำข้อตกลงความร่วมมือกับมหาวิทยาลัยโตเกียว ประเทศญี่ปุ่น และมหาวิทยาลัยมหิดล ในโครงการวิจัยวัณโรคระดับนานาชาติ พัฒนาเทคโนโลยีนวัตกรรมใหม่ทางวิทยาศาสตร์การแพทย์ สำหรับตรวจหาลักษณะทางพันธุกรรมของมนุษย์และเชื้อวัณโรค ช่วยให้วินิจฉัยวัณโรคได้แม่นยำ รวดเร็วขึ้น สามารถเลือกใช้ยาและปรับขนาดยาต้านวัณโรคให้เหมาะสม ลดอาการไม่พึงประสงค์จากยาต้านวัณโรค เพื่อมุ่งสู่นโยบายยุติวัณโรค	กรมวิทยาศาสตร์การแพทย์ ประกาศ ความสำเร็จ การ ทำ ข้อตกลง ความ ร่วมมือ กับ มหาวิทยาลัย โตเกียว ประเทศ ญี่ปุ่น และ มหาวิทยาลัยมหิดล ใน โครงการวิจัย วัณโรค ระดับนานาชาติ พัฒนา เทคโนโลยี นวัตกรรม ใหม่ ทาง วิทยาศาสตร์ การแพทย์ สำหรับ ตรวจหา ลักษณะทางพันธุกรรม ของ มนุษย์ และ เชื้อ วัณโรค ช่วย ให้ วินิจฉัย วัณโรค ได้ แม่นยำ รวดเร็ว ขึ้น สามารถ เลือก ใช้ ยา และ ปรับ ขนาดยา ต้าน วัณโรค ให้ เหมาะสม ลด อาการ ไม่ พึงประสงค์ จาก ยา ต้าน วัณโรค เพื่อ มุ่ง สู่ นโยบาย ยุติ วัณโรค	ข่าวจริง
⋮	⋮	⋮	⋮
3195	หมอเตือน "ไข้หวัดใหญ่" ระบาดหนักแน่ หลังป่วยสูงตั้งแต่ต้นปีรวมกว่า 1.52 แสนราย ตาย 10 ราย คาดป่วย 2 แสนรายเป็นอย่างต่ำ แต่ตัวเลขจริงอาจถึงล้านคน ห่วงฤดูฝน-เปิดเทอมยิ่งเพิ่มการระบาด	หมอ เตือน " ไข้หวัดใหญ่ " ระบาด หนัก แน่ หลัง ป่วย สูง ตั้งแต่ ต้นปี รวม กว่า 1.52 แสน ราย ตาย 10 ราย คาด ป่วย 2 แสน ราย เป็น อย่าง ต่ำ แต่ ตัวเลข จริง อาจ ถึง ล้าน คน ห่วง ฤดูฝน - เปิดเทอม ยิ่ง เพิ่ม การ ระบาด	ข่าวจริง

Table 10 Text Preprocessing Results for Case 1

No.	Topic	Topic_After	Label
3196	ในช่วงการระบาดของโควิด-19 การรักษาสุขอนามัยถือเป็นสิ่งสำคัญ โดยเฉพาะอาหารและน้ำ ที่ต้องสะอาดและปลอดภัยต่อการบริโภค ขณะที่ผู้ประกอบการและผู้สัมผัสอาหาร ควรเรียนรู้หลักสุขาภิบาลอาหาร	ใน ช่วง การ ระบาด ของ โควิด - 19 การรักษา สุขอนามัย ถือเป็น สิ่ง สำคัญ โดยเฉพาะ อาหาร และ น้ำ ที่ ต้อง สะอาด และ ปลอดภัย ต่อ การ บริโภค ขณะที่ ผู้ประกอบการ และ ผู้ สัมผัส อาหาร ควร เรียนรู้ หลัก สุขาภิบาล อาหาร	ข่าว จริง
3197	กรมอนามัย ส่งเสริมหญิงตั้งครรภ์ใช้สมุดบันทึกสุขภาพแม่และเด็ก เล่มสีชมพู เป็นเครื่องมือในการดูแลสุขภาพตนเองและลูกในครรภ์ ช่วงตั้งครรภ์อย่างต่อเนื่องไปจนเด็กเข้าโรงเรียน	กรมอนามัย ส่งเสริม หญิง ตั้งครรภ์ ใช้ สมุดบันทึก สุขภาพ แม่ และ เด็ก เล่ม สีชมพู เป็น เครื่องมือ ใน การ ดูแล สุขภาพ ตนเอง และ ลูก ใน ครรภ์ ช่วง ตั้งครรภ์ อย่าง ต่อเนื่อง ไป จน เด็ก เข้า โรงเรียน	ข่าว จริง
3198	ซื้อ-ขายใบกระท่อม ไม่ต้องขออนุญาตกับทางอย.	ซื้อ - ขาย ใบ กระท่อม ไม่ต้อง ขอ อนุญาต กับ ทาง อย.	ข่าว จริง
3199	หากรพ.ขอรับบริจาคอุปกรณ์ทางการแพทย์ให้แจ้งมาที่องค์การเภสัชกรรม	หาก รพ. ขอ รับบริจาค อุปกรณ์ ทาง การแพทย์ ให้ แจ้ง มา ที่ องค์การ เภสัชกรรม	ข่าว ปลอม

4.1.2 Text Preprocessing Results for Case 2

This case involves cleaning the data using only the Replacing Numbers with Thai Words method out of the three text preprocessing methods. Examples of the data before and after cleaning, as specified, are shown in Table 11.

Table 11 Text Preprocessing Results for Case 2

No.	Topic	Topic_After	Label
0	กรมควบคุมโรค กระทรวงสาธารณสุข ลงพื้นที่สนามบินดอนเมืองตรวจเยี่ยมประชาสัมพันธ์ และบังคับใช้กฎหมายสถานที่สาธารณะปลอดบุหรี่ พบนักท่องเที่ยว และประชาชนผู้ใช้บริการให้การตอบรับและยินยอมปฏิบัติตามกฎหมายสถานที่สาธารณะปลอดบุหรี่ฉบับใหม่ ซึ่งมีผลตั้งแต่วันที่ 3 กุมภาพันธ์ 2562 ที่ผ่านมา	กรม ควบคุม โรค กระทรวง สาธารณสุข ลง พื้นที่ สนามบิน ดอน เมือง ตรวจ เยี่ยม ประชาสัมพันธ์ และ บังคับใช้ กฎหมาย สถานที่สาธารณะ ปลอด บุหรี่ พบ นักท่องเที่ยว และ ประชาชน ผู้ใช้บริการ ให้ การตอบรับ และ ยินยอม ปฏิบัติ ตามกฎหมาย สถานที่สาธารณะ ปลอด บุหรี่ ฉบับ ใหม่ ซึ่ง มีผล ตั้งแต่ วันที่ สาม กุมภาพันธ์ สอง พัน ห้า ร้อย หก สิบ สอง ที่ผ่านมา	ข่าวจริง
1	ยาป้องกันและรักษาโรคไวรัสปอดอักเสบ COVID-19	ยา ป้องกัน และ รักษาโรค ไวรัส ปอด อักเสบ COVID- สิบ เก้า	ข่าวปลอม
2	สมุนไพรพิลังกาสา รักษาโรคมะเร็ง	สมุนไพร พิลังกาสา รักษา โรคมะเร็ง	ข่าวปลอม
3	ไข้เลือดออก ระบาด "ชัยภูมิ-ระยอง-ขอนแก่นแม่ฮ่องสอน-นครราชสีมา" อัตราป่วยสูงสุด ทั่วไทยพบป่วย 2.5 หมื่นคน เสียชีวิต 15 ราย ปัจจุบันเพิ่มความเสี่ยง หลัง อุตุฯพยากรณ์ ฝนจะตกหลายพื้นที่ งัดมาตรการ 3 เก็บออกมารับมือ	ไข้เลือดออก ระบาด " ชัยภูมิ - ระยอง - ขอนแก่น แม่ฮ่องสอน - นครราชสีมา " อัตรา ป่วย สูงสุด ทั่ว ไทย พบ ป่วย สอง . ห้า หมื่น คน เสียชีวิต สิบห้า ราย ปัจจุบัน เพิ่ม ความเสี่ยง หลัง อุตุ ฯ พยากรณ์ ฝน จะ ตก หลาย พื้นที่ งัด มาตรการ สาม เก็บ ออกมา รับมือ	ข่าวจริง

Table 11 Text Preprocessing Results for Case 2

No.	Topic	Topic_After	Label
4	กรมวิทยาศาสตร์การแพทย์ประกาศความสำเร็จการทำข้อตกลงความร่วมมือกับมหาวิทยาลัยโตเกียว ประเทศญี่ปุ่น และมหาวิทยาลัยมหิดล ในโครงการวิจัยวัณโรคระดับนานาชาติ พัฒนาเทคโนโลยีนวัตกรรมใหม่ทางวิทยาศาสตร์การแพทย์ สำหรับตรวจหาลักษณะทางพันธุกรรมของมนุษย์และเชื้อวัณโรค ช่วยให้วินิจฉัยวัณโรคได้แม่นยำ รวดเร็วขึ้น สามารถเลือกใช้ยาและปรับขนาดยาต้านวัณโรคให้เหมาะสม ลดอาการไม่พึงประสงค์จากยาต้านวัณโรค เพื่อมุ่งสู่นโยบายยุติวัณโรค	กรมวิทยาศาสตร์การแพทย์ ประกาศ ความสำเร็จ การ ทำ ข้อตกลง ความ ร่วมมือ กับ มหาวิทยาลัย โตเกียว ประเทศ ญี่ปุ่น และ มหาวิทยาลัยมหิดล ใน โครงการวิจัย วัณโรค ระดับนานาชาติ พัฒนา เทคโนโลยี นวัตกรรม ใหม่ ทาง วิทยาศาสตร์ การแพทย์ สำหรับ ตรวจหา ลักษณะทางพันธุกรรม ของ มนุษย์ และ เชื้อ วัณโรค ช่วย ให้ วินิจฉัย วัณโรค ได้ แม่นยำ รวดเร็ว ขึ้น สามารถ เลือก ใช้ ยา และ ปรับ ขนาดยา ต้าน วัณโรค ให้ เหมาะสม ลด อาการ ไม่ พึงประสงค์ จาก ยา ต้าน วัณโรค เพื่อ มุ่ง สู่ นโยบาย ยุติ วัณโรค	ข่าวจริง
⋮	⋮	⋮	⋮
3195	หมอเตือน "ไข้หวัดใหญ่" ระบาดหนักแน่ หลังป่วยสูงตั้งแต่ต้นปีรวมกว่า 1.52 แสนราย ตาย 10 ราย คาดป่วย 2 แสนรายเป็นอย่างต่ำ แต่ตัวเลขจริงอาจถึงล้านคน ห่วงฤดูฝน-เปิดเทอมยิ่งเพิ่มการระบาด	หมอ เตือน " ไข้หวัดใหญ่ " ระบาด หนัก แน่ หลัง ป่วย สูง ตั้งแต่ ต้นปี รวม กว่า หนึ่ง . ห้า สิบสอง แสน ราย ตาย สิบ ราย คาด ป่วย สอง แสน ราย เป็น อย่าง ต่ำ แต่ ตัวเลข จริง อาจ ถึง ล้าน คน ห่วง ฤดูฝน - เปิดเทอม ยิ่ง เพิ่ม การ ระบาด	ข่าวจริง

Table 11 Text Preprocessing Results for Case 2

No.	Topic	Topic_After	Label
3196	ในช่วงการระบาดของโควิด-19 การรักษาสุขอนามัยถือเป็นสิ่งสำคัญ โดยเฉพาะอาหารและน้ำ ที่ต้องสะอาดและปลอดภัยต่อการบริโภค ขณะที่ผู้ประกอบการและผู้สัมผัสอาหาร ควรเรียนรู้หลักสุขาภิบาลอาหาร	ใน ช่วง การ ระบาด ของ โควิด - สิบ เก้า การรักษา สุขอนามัย ถือเป็น สิ่ง สำคัญ โดยเฉพาะ อาหาร และ น้ำ ที่ ต้อง สะอาด และ ปลอดภัย ต่อ การ บริโภค ขณะที่ ผู้ประกอบการ และ ผู้ สัมผัส อาหาร ควร เรียนรู้ หลัก สุขาภิบาล อาหาร	ข่าว จริง
3197	กรมอนามัย ส่งเสริมหญิงตั้งครรภ์ใช้สมุดบันทึกสุขภาพแม่และเด็ก เล่มสีชมพู เป็นเครื่องมือในการดูแลสุขภาพตนเองและลูกในครรภ์ ช่วงตั้งครรภ์อย่างต่อเนื่องไปจนเด็กเข้าโรงเรียน	กรมอนามัย ส่งเสริม หญิง ตั้งครรภ์ ใช้ สมุดบันทึก สุขภาพ แม่ และ เด็ก เล่ม สีชมพู เป็น เครื่องมือ ใน การ ดูแล สุขภาพ ตนเอง และ ลูก ใน ครรภ์ ช่วง ตั้งครรภ์ อย่าง ต่อเนื่อง ไป จน เด็ก เข้า โรงเรียน	ข่าว จริง
3198	ซื้อ-ขายใบกระท่อม ไม่ต้องขออนุญาตกับทางอย.	ซื้อ - ขาย ใบ กระท่อม ไม่ต้อง ขอ อนุญาต กับ ทาง อย.	ข่าว จริง
3199	หากรพ.ขอรับบริจาคอุปกรณ์ทางการแพทย์ให้แจ้งมาที่องค์การเภสัชกรรม	หาก รพ. ขอ รับบริจาค อุปกรณ์ ทาง การแพทย์ ให้ แจ้ง มา ที่ องค์การ เภสัชกรรม	ข่าว ปลอม

4.1.3 Text Preprocessing Results for Case 3

This case involves cleaning the data using only the Removing Punctuation method out of the three text preprocessing methods. Examples of the data before and after cleaning, as specified, are shown in Table 12.

Table 12 Text Preprocessing Results for Case 3

No.	Topic	Topic_After	Label
0	กรมควบคุมโรค กระทรวงสาธารณสุข ลงพื้นที่สนามบินดอนเมืองตรวจเยี่ยม ประชาสัมพันธ์ และบังคับใช้กฎหมาย สถานที่สาธารณะปลอดบุหรี่ พบ นักท่องเที่ยว และประชาชนผู้ใช้บริการ ให้การตอบรับและยินยอมปฏิบัติตาม กฎหมายสถานที่สาธารณะปลอดบุหรี่ ฉบับใหม่ ซึ่งมีผลตั้งแต่วันที่ 3 กุมภาพันธ์ 2562 ที่ผ่านมา	กรม ควบคุม โรค กระทรวง สาธารณสุข ลง พื้นที่ สนามบิน ดอน เมือง ตรวจ เยี่ยม ประชาสัมพันธ์ และ บังคับใช้ กฎหมาย สถานที่สาธารณะ ปลอด บุหรี่ พบ นักท่องเที่ยว และ ประชาชน ผู้ใช้บริการ ให้ การตอบรับ และ ยินยอม ปฏิบัติ ตามกฎหมาย สถานที่สาธารณะ ปลอด บุหรี่ ฉบับ ใหม่ ซึ่ง มีผล ตั้งแต่ วันที่ 3 กุมภาพันธ์ 2562 ที่ผ่านมา	ข่าว จริง
1	ยาป้องกันและรักษาโรคไวรัสปอด อักเสบ COVID-19	ยา ป้องกัน และ รักษาโรค ไวรัส ปอด อักเสบ COVID 19	ข่าว ปลอม
2	สมุนไพรพิลังกาสา รักษาโรคมะเร็ง	สมุนไพร พิลังกาสา รักษา โรคมะเร็ง	ข่าว ปลอม
3	ไข้เลือดออก ระบาด "ชัยภูมิ-ระยอง-ขอนแก่นแม่ฮ่องสอน-นครราชสีมา" อัตราป่วยสูงสุด ทั่วไทยพบป่วย 2.5 หมื่นคน เสียชีวิต 15 ราย ปัจจุบันเพิ่ม ความเสี่ยง หลัง อุตุฯพยากรณ์ ฝนจะ ตกหลายพื้นที่ งัดมาตรการ 3 เก็บ ออกมารับมือ	ไข้เลือดออก ระบาด ชัยภูมิ ระยอง ขอนแก่น แม่ฮ่องสอน นครราชสีมา อัตรา ป่วย สูงสุด ทั่ว ไทย พบ ป่วย 25 หมื่น คน เสียชีวิต 15 ราย ปัจจุบัน เพิ่ม ความเสี่ยง หลัง อุตุ พยากรณ์ ฝน จะ ตก หลาย พื้นที่ งัด มาตรการ 3 เก็บ ออกมา รับมือ	ข่าว จริง

Table 12 Text Preprocessing Results for Case 3

No.	Topic	Topic_After	Label
4	กรมวิทยาศาสตร์การแพทย์ประกาศความสำเร็จการทำข้อตกลงความร่วมมือกับมหาวิทยาลัยโตเกียว ประเทศญี่ปุ่น และมหาวิทยาลัยมหิดล ในโครงการวิจัยวัณโรคระดับนานาชาติ พัฒนาเทคโนโลยีนวัตกรรมใหม่ทางวิทยาศาสตร์การแพทย์ สำหรับตรวจหาลักษณะทางพันธุกรรมของมนุษย์และเชื้อวัณโรค ช่วยให้วินิจฉัยวัณโรคได้แม่นยำ รวดเร็วขึ้น สามารถเลือกใช้ยาและปรับขนาดยาต้านวัณโรคให้เหมาะสม ลดอาการไม่พึงประสงค์จากยาต้านวัณโรค เพื่อมุ่งสู่นโยบายยุติวัณโรค	กรมวิทยาศาสตร์การแพทย์ ประกาศ ความสำเร็จ การ ทำ ข้อตกลง ความ ร่วมมือ กับ มหาวิทยาลัย โตเกียว ประเทศ ญี่ปุ่น และ มหาวิทยาลัยมหิดล ใน โครงการวิจัย วัณโรค ระดับนานาชาติ พัฒนา เทคโนโลยี นวัตกรรม ใหม่ ทาง วิทยาศาสตร์ การแพทย์ สำหรับ ตรวจหา ลักษณะทางพันธุกรรม ของ มนุษย์ และ เชื้อ วัณโรค ช่วย ให้ วินิจฉัย วัณโรค ได้ แม่นยำ รวดเร็ว ขึ้น สามารถ เลือก ใช้ ยา และ ปรับ ขนาดยา ต้าน วัณโรค ให้ เหมาะสม ลด อาการ ไม่ พึงประสงค์ จาก ยา ต้าน วัณโรค เพื่อ มุ่ง สู่ นโยบาย ยุติ วัณโรค	ข่าวจริง
⋮	⋮	⋮	⋮
3195	หมอเตือน "ไข้หวัดใหญ่" ระบาดหนักแน่ หลังป่วยสูงตั้งแต่ต้นปีรวมกว่า 1.52 แสนราย ตาย 10 ราย คาดป่วย 2 แสนรายเป็นอย่างต่ำ แต่ตัวเลขจริงอาจถึงล้านคน ห่วงฤดูฝน-เปิดเทอมยิ่งเพิ่มการระบาด	หมอ เตือน ไข้หวัดใหญ่ ระบาด หนัก แน่ หลัง ป่วย สูง ตั้งแต่ ต้นปี รวม กว่า 152 แสน ราย ตาย 10 ราย คาด ป่วย 2 แสน ราย เป็น อย่าง ต่ำ แต่ ตัวเลข จริง อาจ ถึง ล้าน คน ห่วง ฤดู ฝน เปิดเทอม ยิ่ง เพิ่ม การ ระบาด	ข่าวจริง

Table 12 Text Preprocessing Results for Case 3

No.	Topic	Topic_After	Label
3196	ในช่วงการระบาดของโควิด-19 การรักษาสุขอนามัยถือเป็นสิ่งสำคัญ โดยเฉพาะอาหารและน้ำ ที่ต้องสะอาดและปลอดภัยต่อการบริโภค ขณะที่ผู้ประกอบการและผู้สัมผัสอาหาร ควรเรียนรู้หลักสุขาภิบาลอาหาร	ใน ช่วง การ ระบาด ของ โควิด 19 การรักษา สุขอนามัย ถือเป็น สิ่ง สำคัญ โดยเฉพาะ อาหาร และ น้ำ ที่ ต้อง สะอาด และ ปลอดภัย ต่อ การ บริโภค ขณะที่ ผู้ประกอบการ และ ผู้ สัมผัส อาหาร ควร เรียนรู้ หลัก สุขาภิบาล อาหาร	ข่าว จริง
3197	กรมอนามัย ส่งเสริมหญิงตั้งครรภ์ใช้สมุดบันทึกสุขภาพแม่และเด็ก เล่มสีชมพู เป็นเครื่องมือในการดูแลสุขภาพตนเองและลูกในครรภ์ ช่วงตั้งครรภ์อย่างต่อเนื่องไปจนเด็กเข้าโรงเรียน	กรมอนามัย ส่งเสริม หญิง ตั้งครรภ์ ใช้ สมุดบันทึก สุขภาพ แม่ และ เด็ก เล่ม สีชมพู เป็น เครื่องมือ ใน การ ดูแล สุขภาพ ตนเอง และ ลูก ใน ครรภ์ ช่วง ตั้งครรภ์ อย่าง ต่อเนื่อง ไป จน เด็ก เข้า โรงเรียน	ข่าว จริง
3198	ซื้อ-ขายใบกระท่อม ไม่ต้องขออนุญาตกับทางอย.	ซื้อขาย ใบ กระท่อม ไม่ต้อง ขอ อนุญาต กับ ทา งอย	ข่าว จริง
3199	หากรพ.ขอรับบริจาคอุปกรณ์ทางการแพทย์ให้แจ้งมาที่องค์การเภสัชกรรม	หาก รพ ขอ รับบริจาค อุปกรณ์ ทาง การแพทย์ ให้ แจ้ง มา ที่ องค์การ เภสัชกรรม	ข่าว ปลอม

4.1.4 Text Preprocessing Results for Case 4

This case involves cleaning the data using only the Removing Stop Words method out of the three text preprocessing methods. Examples of the data before and after cleaning, as specified, are shown in Table 13.

Table 13 Text Preprocessing Results for Case 4

No.	Topic	Topic_After	Label
0	กรมควบคุมโรค กระทรวงสาธารณสุข ลงพื้นที่สนามบินดอนเมืองตรวจเยี่ยม ประชาสัมพันธ์ และบังคับใช้กฎหมาย สถานที่สาธารณะปลอดบุหรี่ พบ นักท่องเที่ยว และประชาชนผู้ใช้บริการ ให้การตอบรับและยินยอมปฏิบัติตาม กฎหมายสถานที่สาธารณะปลอดบุหรี่ ฉบับใหม่ ซึ่งมีผลตั้งแต่วันที่ 3 กุมภาพันธ์ 2562 ที่ผ่านมา	กรม ควบคุม โรค กระทรวง สาธารณสุข พื้นที่ สนามบิน ดอนเมือง ตรวจ เยี่ยม ประชาสัมพันธ์ บังคับใช้ กฎหมาย สถานที่สาธารณะ ปลอด บุหรี่ นักท่องเที่ยว ประชาชน ผู้ใช้บริการ การตอบรับ ยินยอม ตาม กฎหมาย สถานที่สาธารณะ ปลอด บุหรี่ ฉบับ มีผล วันที่ 3 กุมภาพันธ์ 2562 ที่ผ่านมา	ข่าว จริง
1	ยาป้องกันและรักษาโรคไวรัสปอด อักเสบ COVID-19	ยา ป้องกัน รักษาโรค ไวรัส ปอด อักเสบ COVID- 19	ข่าว ปลอม
2	สมุนไพรพิลังกาสา รักษาโรคมะเร็ง	สมุนไพร พิลังกาสา รักษา โรคมะเร็ง	ข่าว ปลอม
3	ไข้เลือดออก ระบาด "ชัยภูมิ-ระยอง-ขอนแก่นแม่ฮ่องสอน-นครราชสีมา" อัตราป่วยสูงสุด ทั่วไทยพบป่วย 2.5 หมื่นคน เสียชีวิต 15 ราย ปัจจุบันเพิ่ม ความเสี่ยง หลัง อุตุฯพยากรณ์ ฝนจะ ตกหลายพื้นที่ งัดมาตรการ 3 เก็บ ออกมารับมือ	ไข้เลือดออก ระบาด " ชัยภูมิ - ระยอง - ขอนแก่น แม่ฮ่องสอน - นครราชสีมา " อัตรา ป่วย ไทย ป่วย 2.5 หมื่น คน เสียชีวิต 15 ความเสี่ยง อุตุฯ พยากรณ์ ฝน ตก พื้นที่ งัด มาตรการ 3 ออกมา รับมือ	ข่าว จริง

Table 13 Text Preprocessing Results for Case 4

No.	Topic	Topic_After	Label
4	กรมวิทยาศาสตร์การแพทย์ประกาศความสำเร็จการทำข้อตกลงความร่วมมือกับมหาวิทยาลัยโตเกียว ประเทศญี่ปุ่น และมหาวิทยาลัยมหิดล ในโครงการวิจัยวัณโรคระดับนานาชาติ พัฒนาเทคโนโลยีนวัตกรรมใหม่ทางวิทยาศาสตร์การแพทย์ สำหรับตรวจหาลักษณะทางพันธุกรรมของมนุษย์และเชื้อวัณโรค ช่วยให้วินิจฉัยวัณโรคได้แม่นยำ รวดเร็วขึ้น สามารถเลือกใช้ยาและปรับขนาดยาต้านวัณโรคให้เหมาะสม ลดอาการไม่พึงประสงค์จากยาต้านวัณโรค เพื่อมุ่งสู่นโยบายยุติวัณโรค	กรมวิทยาศาสตร์การแพทย์ ประกาศ ความสำเร็จ ทำ ข้อตกลง ความร่วมมือ มหาวิทยาลัย โตเกียว ประเทศ ญี่ปุ่น มหาวิทยาลัยมหิดล โครงการวิจัย วัณโรค ระดับนานาชาติ พัฒนา เทคโนโลยี นวัตกรรม ทางวิทยาศาสตร์การแพทย์ สำหรับ ตรวจหา ลักษณะทางพันธุกรรม มนุษย์ เชื้อ วัณโรค วินิจฉัย วัณโรค แม่นยำ เลือก ยา ขนาดยา ต้าน วัณโรค เหมาะสม ลดอาการ พึงประสงค์ ยา ต้าน วัณโรค นโยบาย ยุติ วัณโรค	ข่าวจริง
⋮	⋮	⋮	⋮
3195	หมอเตือน "ไข้หวัดใหญ่" ระบาดหนักแน่ หลังป่วยสูงตั้งแต่ต้นปีรวมกว่า 1.52 แสนราย ตาย 10 ราย คาดป่วย 2 แสนรายเป็นอย่างต่ำ แต่ตัวเลขจริงอาจถึงล้านคน ห่วงฤดูฝน-เปิดเทอมยิ่งเพิ่มการระบาด	หมอ เตือน " ไข้หวัดใหญ่ " ระบาด หนัก แน่ ป่วย ต้นปี 1.52 แสน ตาย 10 คาด ป่วย 2 แสน ต่ำ ตัวเลข ล้าน คน ห่วง ฤดูฝน - เปิดเทอม ระบาด	ข่าวจริง
3196	ในช่วงการระบาดของโควิด-19 การรักษาสุขอนามัยถือเป็นสิ่งสำคัญ โดยเฉพาะอาหารและน้ำ ที่ต้องสะอาดและปลอดภัยต่อการบริโภค ขณะที่ผู้ประกอบการและผู้สัมผัสอาหาร ควรเรียนรู้หลักสุขาภิบาลอาหาร	ระบาด โควิด - 19 การรักษา สุขอนามัย ถือเป็น โดยเฉพาะ อาหาร น้ำ สะอาด ปลอดภัย บริโภค ผู้ประกอบการ สัมผัส อาหาร เรียนรู้ หลัก สุขาภิบาล อาหาร	ข่าวจริง

Table 13 Text Preprocessing Results for Case 4

No.	Topic	Topic_After	Label
3197	กรมอนามัย ส่งเสริมหญิงตั้งครรภ์ใช้สมุดบันทึกสุขภาพแม่และเด็ก เล่มสีชมพู เป็นเครื่องมือในการดูแลสุขภาพตนเองและลูกในครรภ์ ช่วงตั้งครรภ์อย่างต่อเนื่องไปจนเด็กเข้าโรงเรียน	กรมอนามัย ส่งเสริม หญิง ตั้งครรภ์ สมุดบันทึก สุขภาพ แม่ เด็ก เล่ม สีชมพู เครื่องมือ ดูแล สุขภาพ ลูก ครรภ์ ตั้งครรภ์ ต่อเนื่อง เด็ก โรงเรียน	ข่าวจริง
3198	ซื้อ-ขายใบกระท่อม ไม่ต้องขออนุญาตกับทางอย.	ซื้อ - ขาย ใบ กระท่อม ไม่ต้อง ขอ อนุญาต อย.	ข่าวจริง
3199	หากรพ.ขอรับบริจาคอุปกรณ์ทางการแพทย์ให้แจ้งมาที่องค์การเภสัชกรรม	รพ. รับบริจาค อุปกรณ์ ทางการแพทย์ แจ้ง องค์การเภสัชกรรม	ข่าวปลอม

4.1.5 Text Preprocessing Results for Case 5

This case involves cleaning the data using the Replacing Numbers with Thai Words and Removing Punctuation methods out of the three text preprocessing methods. Examples of the data before and after cleaning, as specified, are shown in Table 14.

Table 14 Text Preprocessing Results for Case 5

No.	Topic	Topic_After	Label
0	กรมควบคุมโรค กระทรวงสาธารณสุข ลงพื้นที่สนามบินดอนเมืองตรวจเยี่ยมประชาสัมพันธ์ และบังคับใช้กฎหมายสถานที่สาธารณะปลอดบุหรี่ พบนักท่องเที่ยว และประชาชนผู้ใช้บริการให้การตอบรับและยินยอมปฏิบัติตามกฎหมายสถานที่สาธารณะปลอดบุหรี่ฉบับใหม่ ซึ่งมีผลตั้งแต่วันที่ 3 กุมภาพันธ์ 2562 ที่ผ่านมา	กรม ควบคุม โรค กระทรวง สาธารณสุข ลง พื้นที่ สนามบิน ดอน เมือง ตรวจ เยี่ยม ประชาสัมพันธ์ และ บังคับใช้ กฎหมาย สถานที่สาธารณะ ปลอด บุหรี่ พบ นักท่องเที่ยว และ ประชาชน ผู้ใช้บริการ ให้ การตอบรับ และ ยินยอม ปฏิบัติ ตามกฎหมาย สถานที่สาธารณะ ปลอด บุหรี่ ฉบับ ใหม่ ซึ่ง มีผล ตั้งแต่ วันที่ สาม กุมภาพันธ์ สอง พัน ห้า ร้อย หก สิบ สอง ที่ผ่านมา	ข่าว จริง
1	ยาป้องกันและรักษาโรคไวรัสปอดอักเสบ COVID-19	ยา ป้องกัน และ รักษาโรค ไวรัส ปอด อักเสบ COVID สิบ เก้า	ข่าว ปลอม
2	สมุนไพรพิลังกาสา รักษาโรคมะเร็ง	สมุนไพร พิลังกาสา รักษา โรคมะเร็ง	ข่าว ปลอม
3	ไข้เลือดออก ระบาด "ชัยภูมิ-ระยอง-ขอนแก่นแม่ฮ่องสอน-นครราชสีมา" อัตราป่วยสูงสุด ทั่วไทยพบป่วย 2.5 หมื่นคน เสียชีวิต 15 ราย ปัจจุบันเพิ่มความเสี่ยง หลัง อุตุฯพยากรณ์ ฝนจะตกหลายพื้นที่ งัดมาตรการ 3 เก็บออกมารับมือ	ไข้เลือดออก ระบาด ชัยภูมิ ระยอง ขอนแก่น แม่ฮ่องสอน นครราชสีมา อัตรา ป่วย สูงสุด ทั่ว ไทย พบ ป่วย สอง ห้า หมื่น คน เสียชีวิต สิบห้า ราย ปัจจุบัน เพิ่ม ความเสี่ยง หลัง อุตุ พยากรณ์ ฝน จะ ตก หลาย พื้นที่ งัด มาตรการ สาม เก็บ ออกมา รับมือ	ข่าว จริง

Table 14 Text Preprocessing Results for Case 5

No.	Topic	Topic_After	Label
4	กรมวิทยาศาสตร์การแพทย์ประกาศความสำเร็จการทำข้อตกลงความร่วมมือกับมหาวิทยาลัยโตเกียว ประเทศญี่ปุ่น และมหาวิทยาลัยมหิดล ในโครงการวิจัยวัณโรคระดับนานาชาติ พัฒนาเทคโนโลยีนวัตกรรมใหม่ทางวิทยาศาสตร์การแพทย์ สำหรับตรวจหาลักษณะทางพันธุกรรมของมนุษย์และเชื้อวัณโรค ช่วยให้วินิจฉัยวัณโรคได้แม่นยำ รวดเร็วขึ้น สามารถเลือกใช้ยาและปรับขนาดยาต้านวัณโรคให้เหมาะสม ลดอาการไม่พึงประสงค์จากยาต้านวัณโรค เพื่อมุ่งสู่นโยบายยุติวัณโรค	กรมวิทยาศาสตร์การแพทย์ ประกาศ ความสำเร็จ การ ทำ ข้อตกลง ความ ร่วมมือ กับ มหาวิทยาลัย โตเกียว ประเทศ ญี่ปุ่น และ มหาวิทยาลัยมหิดล ใน โครงการวิจัย วัณโรค ระดับนานาชาติ พัฒนา เทคโนโลยี นวัตกรรม ใหม่ ทาง วิทยาศาสตร์ การแพทย์ สำหรับ ตรวจหา ลักษณะทางพันธุกรรม ของ มนุษย์ และ เชื้อ วัณโรค ช่วย ให้ วินิจฉัย วัณโรค ได้ แม่นยำ รวดเร็ว ขึ้น สามารถ เลือก ใช้ ยา และ ปรับ ขนาดยา ต้าน วัณโรค ให้ เหมาะสม ลด อาการ ไม่ พึงประสงค์ จาก ยา ต้าน วัณโรค เพื่อ มุ่ง สู่ นโยบาย ยุติ วัณโรค	ข่าวจริง
⋮	⋮	⋮	⋮
3195	หมอเตือน "ไข้หวัดใหญ่" ระบาดหนักแน่ หลังป่วยสูงตั้งแต่ต้นปีรวมกว่า 1.52 แสนราย ตาย 10 ราย คาดป่วย 2 แสนรายเป็นอย่างต่ำ แต่ตัวเลขจริงอาจถึงล้านคน ห่วงฤดูฝน-เปิดเทอมยิ่งเพิ่มการระบาด	หมอ เตือน ไข้หวัดใหญ่ ระบาด หนัก แน่ หลัง ป่วย สูง ตั้งแต่ ต้นปี รวม กว่า หนึ่ง ห้า สิบสอง แสน ราย ตาย สิบ ราย คาด ป่วย สอง แสน ราย เป็น อย่าง ต่ำ แต่ ตัวเลข จริง อาจ ถึง ล้าน คน ห่วง ฤดูฝน เปิดเทอม ยิ่ง เพิ่ม การ ระบาด	ข่าวจริง

Table 14 Text Preprocessing Results for Case 5

No.	Topic	Topic_After	Label
3196	ในช่วงการระบาดของโควิด-19 การรักษาสุขอนามัยถือเป็นสิ่งสำคัญ โดยเฉพาะอาหารและน้ำ ที่ต้องสะอาดและปลอดภัยต่อการบริโภค ขณะที่ผู้ประกอบการและผู้สัมผัสอาหาร ควรเรียนรู้หลักสุขาภิบาลอาหาร	ใน ช่วง การ ระบาด ของ โควิด สิบ เก้า การรักษา สุขอนามัย ถือเป็น สิ่ง สำคัญ โดยเฉพาะ อาหาร และ น้ำ ที่ ต้อง สะอาด และ ปลอดภัย ต่อ การ บริโภค ขณะที่ ผู้ประกอบการ และ ผู้ สัมผัส อาหาร ควร เรียนรู้ หลัก สุขาภิบาล อาหาร	ข่าว จริง
3197	กรมอนามัย ส่งเสริมหญิงตั้งครรภ์ใช้สมุดบันทึกสุขภาพแม่และเด็ก เล่มสีชมพู เป็นเครื่องมือในการดูแลสุขภาพตนเองและลูกในครรภ์ ช่วงตั้งครรภ์อย่างต่อเนื่องไปจนเด็กเข้าโรงเรียน	กรมอนามัย ส่งเสริม หญิง ตั้งครรภ์ ใช้ สมุดบันทึก สุขภาพ แม่ และ เด็ก เล่ม สีชมพู เป็น เครื่องมือ ใน การ ดูแล สุขภาพ ตนเอง และ ลูก ใน ครรภ์ ช่วง ตั้งครรภ์ อย่าง ต่อเนื่อง ไป จน เด็ก เข้า โรงเรียน	ข่าว จริง
3198	ซื้อ-ขายใบกระท่อม ไม่ต้องขออนุญาตกับทางอย.	ซื้อขาย ใบ กระท่อม ไม่ต้อง ขอ อนุญาต กับ ทา งอย	ข่าว จริง
3199	หากรพ.ขอรับบริจาคอุปกรณ์ทางการแพทย์ให้แจ้งมาที่องค์การเภสัชกรรม	หาก รพ ขอ รับบริจาค อุปกรณ์ ทาง การแพทย์ ให้ แจ้ง มา ที่ องค์การ เภสัชกรรม	ข่าว ปลอม

4.1.6 Text Preprocessing Results for Case 6

This case involves cleaning the data using the Replacing Numbers with Thai Words and Removing Stop Words methods out of the three text preprocessing methods. Examples of the data before and after cleaning, as specified, are shown in Table 15.

Table 15 Text Preprocessing Results for Case 6

No.	Topic	Topic_After	Label
0	กรมควบคุมโรค กระทรวงสาธารณสุข ลงพื้นที่สนามบินดอนเมืองตรวจเยี่ยมประชาสัมพันธ์ และบังคับใช้กฎหมายสถานที่สาธารณะปลอดบุหรี่ พบนักท่องเที่ยว และประชาชนผู้ใช้บริการให้การตอบรับและยินยอมปฏิบัติตามกฎหมายสถานที่สาธารณะปลอดบุหรี่ฉบับใหม่ ซึ่งมีผลตั้งแต่วันที่ 3 กุมภาพันธ์ 2562 ที่ผ่านมา	กรม ควบคุม โรค กระทรวง สาธารณสุข พื้นที่ สนามบิน ดอนเมือง ตรวจ เยี่ยม ประชาสัมพันธ์ บังคับใช้ กฎหมาย สถานที่สาธารณะ ปลอด บุหรี่ นักท่องเที่ยว ประชาชน ผู้ใช้บริการ การตอบรับ ยินยอม ตาม กฎหมาย สถานที่สาธารณะ ปลอด บุหรี่ ฉบับ มีผล วันที่ สาม กุมภาพันธ์ สอง พัน ห้า ร้อย หก สิบสอง ที่ผ่านมา	ข่าวจริง
1	ยาป้องกันและรักษาโรคไวรัสปอดอักเสบ COVID-19	ยา ป้องกัน รักษาโรค ไวรัส ปอด อักเสบ COVID- สิบ เก้า	ข่าวปลอม
2	สมุนไพรพิลังกาสา รักษาโรคมะเร็ง	สมุนไพร พิลังกาสา รักษา โรคมะเร็ง	ข่าวปลอม
3	ไข้เลือดออก ระบาด "ชัยภูมิ-ระยอง-ขอนแก่นแม่ฮ่องสอน-นครราชสีมา" อัตราป่วยสูงสุด ทั่วไทยพบป่วย 2.5 หมื่นคน เสียชีวิต 15 ราย ปัจจุบันเพิ่มความเสี่ยง หลัง อุตุฯพยากรณ์ ฝนจะตกหลายพื้นที่ งัดมาตรการ 3 เก็บออกมารับมือ	ไข้เลือดออก ระบาด " ชัยภูมิ - ระยอง - ขอนแก่น แม่ฮ่องสอน - นครราชสีมา " อัตรา ป่วย ไทย ป่วย สอง . ห้า หมื่น คน เสียชีวิต สิบห้า ความเสี่ยง อุตุฯ พยากรณ์ ฝน ตก พื้นที่ งัด มาตรการ สาม ออกมา รับมือ	ข่าวจริง

Table 15 Text Preprocessing Results for Case 6

No.	Topic	Topic_After	Label
4	กรมวิทยาศาสตร์การแพทย์ประกาศความสำเร็จการทำข้อตกลงความร่วมมือกับมหาวิทยาลัยโตเกียว ประเทศญี่ปุ่น และมหาวิทยาลัยมหิดล ในโครงการวิจัยวัณโรคระดับนานาชาติ พัฒนาเทคโนโลยีนวัตกรรมใหม่ทางวิทยาศาสตร์การแพทย์ สำหรับตรวจหาลักษณะทางพันธุกรรมของมนุษย์และเชื้อวัณโรค ช่วยให้วินิจฉัยวัณโรคได้แม่นยำ รวดเร็วขึ้น สามารถเลือกใช้ยาและปรับขนาดยาต้านวัณโรคให้เหมาะสม ลดอาการไม่พึงประสงค์จากยาต้านวัณโรค เพื่อมุ่งสู่นโยบายยุติวัณโรค	กรมวิทยาศาสตร์การแพทย์ ประกาศ ความสำเร็จ ทำ ข้อตกลง ความร่วมมือ มหาวิทยาลัย โตเกียว ประเทศ ญี่ปุ่น มหาวิทยาลัยมหิดล โครงการวิจัย วัณโรค ระดับนานาชาติ พัฒนา เทคโนโลยี นวัตกรรม ทางวิทยาศาสตร์การแพทย์ สำหรับ ตรวจหา ลักษณะทางพันธุกรรม มนุษย์ เชื้อ วัณโรค วินิจฉัย วัณโรค แม่นยำ เลือก ยา ขนาดยา ต้าน วัณโรค เหมาะสม ลดอาการ พึงประสงค์ ยา ต้าน วัณโรค นโยบาย ยุติ วัณโรค	ข่าวจริง
⋮	⋮	⋮	⋮
3195	หมอเตือน "ไข้หวัดใหญ่" ระบาดหนักแน่ หลังป่วยสูงตั้งแต่ต้นปีรวมกว่า 1.52 แสนราย ตาย 10 ราย คาดป่วย 2 แสนรายเป็นอย่างต่ำ แต่ตัวเลขจริงอาจถึงล้านคน ห่วงฤดูฝน-เปิดเทอมยิ่งเพิ่มการระบาด	หมอ เตือน " ไข้หวัดใหญ่ " ระบาด หนัก แน่ ป่วย ต้นปี . ห้า สิบสอง แสน ตาย สิบ คาด ป่วย สอง แสน ต่ำ ตัวเลข ล้าน คน ห่วง ฤดูฝน - เปิด เทอม ระบาด	ข่าวจริง
3196	ในช่วงการระบาดของโควิด-19 การรักษาสุขอนามัยถือเป็นสิ่งสำคัญ โดยเฉพาะอาหารและน้ำ ที่ต้องสะอาดและปลอดภัยต่อการบริโภค ขณะที่ผู้ประกอบการและผู้สัมผัสอาหาร ควรเรียนรู้หลักสุขาภิบาลอาหาร	ระบาด โควิด - สิบ เก้า การรักษา สุขอนามัย ถือเป็น โดยเฉพาะ อาหาร น้ำ สะอาด ปลอดภัย บริโภค ผู้ประกอบการ สัมผัส อาหาร เรียนรู้ หลัก สุขาภิบาล อาหาร	ข่าวจริง

Table 15 Text Preprocessing Results for Case 6

No.	Topic	Topic_After	Label
3197	กรมอนามัย ส่งเสริมหญิงตั้งครรภ์ใช้สมุดบันทึกสุขภาพแม่และเด็ก เล่มสีชมพู เป็นเครื่องมือในการดูแลสุขภาพตนเองและลูกในครรภ์ ช่วงตั้งครรภ์อย่างต่อเนื่องไปจนเด็กเข้าโรงเรียน	กรมอนามัย ส่งเสริม หญิง ตั้งครรภ์ สมุดบันทึก สุขภาพ แม่ เด็ก เล่ม สีชมพู เครื่องมือ ดูแล สุขภาพ ลูกครรภ์ ตั้งครรภ์ ต่อเนื่อง เด็ก โรงเรียน	ข่าวจริง
3198	ซื้อ-ขายใบกระท่อม ไม่ต้องขออนุญาตกับทางอย.	ซื้อ - ขาย ใบ กระท่อม ไม่ต้อง ขออนุญาต อย.	ข่าวจริง
3199	หากรพ.ขอรับบริจาคอุปกรณ์ทางการแพทย์ให้แจ้งมาที่องค์การเภสัชกรรม	รพ. รับบริจาค อุปกรณ์ ทางการแพทย์ แจ้ง องค์การเภสัชกรรม	ข่าวปลอม

4.1.7 Text Preprocessing Results for Case 7

This case involves cleaning the data using the Removing Punctuation and Removing Stop Words methods out of the three text preprocessing methods. Examples of the data before and after cleaning, as specified, are shown in Table 16.

Table 16 Text Preprocessing Results for Case 7

No.	Topic	Topic_After	Label
0	กรมควบคุมโรค กระทรวงสาธารณสุข ลงพื้นที่สนามบินดอนเมืองตรวจเยี่ยมประชาสัมพันธ์ และบังคับใช้กฎหมายสถานที่สาธารณะปลอดบุหรี่ พบนักท่องเที่ยว และประชาชนผู้ใช้บริการให้การตอบรับและยินยอมปฏิบัติตามกฎหมายสถานที่สาธารณะปลอดบุหรี่ฉบับใหม่ ซึ่งมีผลตั้งแต่วันที่ 3 กุมภาพันธ์ 2562 ที่ผ่านมา	กรม ควบคุม โรค กระทรวง สาธารณสุข พื้นที่ สนามบิน ดอนเมือง ตรวจ เยี่ยม ประชาสัมพันธ์ บังคับใช้ กฎหมาย สถานที่สาธารณะ ปลอด บุหรี่ นักท่องเที่ยว ประชาชน ผู้ใช้บริการ การตอบรับ ยินยอม ตาม กฎหมาย สถานที่สาธารณะ ปลอด บุหรี่ ฉบับ มีผล วันที่ 3 กุมภาพันธ์ 2562 ที่ผ่านมา	ข่าวจริง

Table 16 Text Preprocessing Results for Case 7

No.	Topic	Topic_After	Label
1	ยาป้องกันและรักษาโรคไวรัสปอดอักเสบ COVID-19	ยา ป้องกัน รักษาโรค ไวรัส ปอดอักเสบ COVID 19	ข่าวปลอม
2	สมุนไพรพิลังกาสา รักษาโรคมะเร็ง	สมุนไพร พิลังกาสา รักษา โรคมะเร็ง	ข่าวปลอม
3	ไข้เลือดออก ระบาด "ชัยภูมิ-ระยอง-ขอนแก่นแม่ฮ่องสอน-นครราชสีมา" อัตราป่วยสูงสุด ทั่วไทยพบป่วย 2.5 หมื่นคน เสียชีวิต 15 ราย ปัจจุบันเพิ่มความเสี่ยง หลัง อุตุฯพยากรณ์ ฝนจะตกหลายพื้นที่ งัดมาตรการ 3 เก็บออกมารับมือ	ไข้เลือดออก ระบาด ชัยภูมิ ระยอง ขอนแก่น แม่ฮ่องสอน นครราชสีมา อัตรา ป่วย ไทย ป่วย 25 หมื่น คน เสียชีวิต 15 ความเสี่ยง อุตุ พยากรณ์ ฝน ตก พื้นที่ งัด มาตรการ 3 ออกมา รับมือ	ข่าวจริง
4	กรมวิทยาศาสตร์การแพทย์ประกาศความสำเร็จการทำข้อตกลงความร่วมมือกับมหาวิทยาลัยโตเกียว ประเทศญี่ปุ่น และมหาวิทยาลัยมหิดล ในโครงการวิจัยวัณโรคระดับนานาชาติ พัฒนาเทคโนโลยีนวัตกรรมใหม่ทางวิทยาศาสตร์การแพทย์ สำหรับตรวจหาลักษณะทางพันธุกรรมของมนุษย์และเชื้อวัณโรค ช่วยให้วินิจฉัยวัณโรคได้แม่นยำ รวดเร็วขึ้น สามารถเลือกใช้ยาและปรับขนาดยาต้านวัณโรคให้เหมาะสม ลดอาการไม่พึงประสงค์จากยาต้านวัณโรค เพื่อมุ่งสู่นโยบายยุติวัณโรค	กรมวิทยาศาสตร์การแพทย์ ประกาศ ความสำเร็จ ทำ ข้อตกลง ความร่วมมือ มหาวิทยาลัย โตเกียว ประเทศ ญี่ปุ่น มหาวิทยาลัยมหิดล โครงการวิจัย วัณโรค ระดับนานาชาติ พัฒนา เทคโนโลยี นวัตกรรม ทางวิทยาศาสตร์การแพทย์ สำหรับ ตรวจหา ลักษณะทางพันธุกรรม มนุษย์ เชื้อ วัณโรค วินิจฉัย วัณโรค แม่นยำ เลือก ยา ขนาดยา ต้าน วัณโรค เหมาะสม ลดอาการ พึงประสงค์ ยา ต้าน วัณโรค นโยบาย ยุติ วัณโรค	ข่าวจริง
⋮	⋮	⋮	⋮

Table 16 Text Preprocessing Results for Case 7

No.	Topic	Topic_After	Label
3195	หมอเตือน "ไข้หวัดใหญ่" ระบาดหนักแน่ หลังป่วยสูงตั้งแต่ต้นปีรวมกว่า 1.52 แสนราย ตาย 10 ราย คาดป่วย 2 แสนรายเป็นอย่างต่ำ แต่ตัวเลขจริงอาจถึงล้านคน ห่วงฤดูฝน-เปิดเทอมยิ่งเพิ่มการระบาด	หมอ เตือน ไข้หวัดใหญ่ ระบาด หนัก แน่ ป่วย ต้นปี 152 แสน ตาย 10 คาด ป่วย 2 แสน ต่ำ ตัวเลข ล้าน คน ห่วง ฤดูฝน เปิดเทอม ระบาด	ข่าวจริง
3196	ในช่วงการระบาดของโควิด-19 การรักษาสุขอนามัยถือเป็นสิ่งสำคัญ โดยเฉพาะอาหารและน้ำ ที่ต้องสะอาดและปลอดภัยต่อการบริโภค ขณะที่ผู้ประกอบการและผู้สัมผัสอาหาร ควรเรียนรู้หลักสุขาภิบาลอาหาร	ระบาด โควิด 19 การรักษา สุขอนามัย ถือเป็น โดยเฉพาะ อาหาร น้ำ สะอาด ปลอดภัย บริโภค ผู้ประกอบการ สัมผัส อาหาร เรียนรู้ หลัก สุขาภิบาล อาหาร	ข่าวจริง
3197	กรมอนามัย ส่งเสริมหญิงตั้งครรภ์ใช้สมุดบันทึกสุขภาพแม่และเด็ก เล่มสีชมพู เป็นเครื่องมือในการดูแลสุขภาพตนเองและลูกในครรภ์ ช่วงตั้งครรภ์อย่างต่อเนื่องไปจนเด็กเข้าโรงเรียน	กรมอนามัย ส่งเสริม หญิง ตั้งครรภ์ สมุดบันทึก สุขภาพ แม่ เด็ก เล่ม สีชมพู เครื่องมือ ดูแล สุขภาพ ลูกครรภ์ ตั้งครรภ์ ต่อเนื่อง เด็ก โรงเรียน	ข่าวจริง
3198	ซื้อ-ขายใบกระท่อม ไม่ต้องขออนุญาตกับทางอย.	ซื้อขาย ใบ กระท่อม ไม่ต้อง ขออนุญาต ทา งอย	ข่าวจริง
3199	หากรพ.ขอรับบริจาคอุปกรณ์ทางการแพทย์ให้แจ้งมาที่องค์การเภสัชกรรม	รพ รับบริจาค อุปกรณ์ ทางการแพทย์ แจ้ง องค์การเภสัชกรรม	ข่าวปลอม

4.1.8 Text Preprocessing Results for Case 8

This case involves cleaning the data using all three text preprocessing methods (Replacing Numbers with Thai Words, Removing Punctuation, Removing Stop Words). Examples of the data before and after cleaning, as specified, are shown in Table 17.

Table 17 Text Preprocessing Results for Case 8

No.	Topic	Topic_After	Label
0	กรมควบคุมโรค กระทรวงสาธารณสุข ลงพื้นที่สนามบินดอนเมืองตรวจเยี่ยมประชาสัมพันธ์ และบังคับใช้กฎหมายสถานที่สาธารณะปลอดบุหรี่ พบนักท่องเที่ยว และประชาชนผู้ใช้บริการให้การตอบรับและยินยอมปฏิบัติตามกฎหมายสถานที่สาธารณะปลอดบุหรี่ฉบับใหม่ ซึ่งมีผลตั้งแต่วันที่ 3 กุมภาพันธ์ 2562 ที่ผ่านมา	กรม ควบคุม โรค กระทรวง สาธารณสุข พื้นที่ สนามบิน ดอนเมือง ตรวจ เยี่ยม ประชาสัมพันธ์ บังคับใช้ กฎหมาย สถานที่สาธารณะ ปลอด บุหรี่ นักท่องเที่ยว ประชาชน ผู้ใช้บริการ การตอบรับ ยินยอม ตาม กฎหมาย สถานที่สาธารณะ ปลอด บุหรี่ ฉบับ มีผล วันที่ สาม กุมภาพันธ์ สอง พัน ห้า ร้อย หก สิบสอง ที่ผ่านมา	ข่าวจริง
1	ยาป้องกันและรักษาโรคไวรัสปอดอักเสบ COVID-19	ยา ป้องกัน รักษาโรค ไวรัส ปอด อักเสบ COVID สิบ เก้า	ข่าวปลอม
2	สมุนไพรพิลังกาสา รักษาโรคมะเร็ง	สมุนไพร พิลังกาสา รักษา โรคมะเร็ง	ข่าวปลอม
3	ไข้เลือดออก ระบาด "ชัยภูมิ-ระยอง-ขอนแก่นแม่ฮ่องสอน-นครราชสีมา" อัตราป่วยสูงสุด ทั่วไทยพบป่วย 2.5 หมื่นคน เสียชีวิต 15 ราย ปัจจุบันเพิ่มความเสี่ยง หลัง อุตุฯพยากรณ์ ฝนจะตกหลายพื้นที่ งัดมาตรการ 3 เก็บออกมารับมือ	ไข้เลือดออก ระบาด ชัยภูมิ ระยอง ขอนแก่น แม่ฮ่องสอน นครราชสีมา อัตรา ป่วย ไทย ป่วย สอง ห้า หมื่น คน เสียชีวิต สิบห้า ความเสี่ยง อุตุ พยากรณ์ ฝน ตก พื้นที่ งัด มาตรการ สาม ออกมา รับมือ	ข่าวจริง

Table 17 Text Preprocessing Results for Case 8

No.	Topic	Topic_After	Label
4	กรมวิทยาศาสตร์การแพทย์ประกาศความสำเร็จการทำข้อตกลงความร่วมมือกับมหาวิทยาลัยโตเกียว ประเทศญี่ปุ่น และมหาวิทยาลัยมหิดล ในโครงการวิจัยวัณโรคระดับนานาชาติ พัฒนาเทคโนโลยีนวัตกรรมใหม่ทางวิทยาศาสตร์การแพทย์ สำหรับตรวจหาลักษณะทางพันธุกรรมของมนุษย์และเชื้อวัณโรค ช่วยให้วินิจฉัยวัณโรคได้แม่นยำ รวดเร็วขึ้น สามารถเลือกใช้ยาและปรับขนาดยาต้านวัณโรคให้เหมาะสม ลดอาการไม่พึงประสงค์จากยาต้านวัณโรค เพื่อมุ่งสู่นโยบายยุติวัณโรค	กรมวิทยาศาสตร์การแพทย์ ประกาศ ความสำเร็จ ทำ ข้อตกลง ความร่วมมือ มหาวิทยาลัย โตเกียว ประเทศ ญี่ปุ่น มหาวิทยาลัยมหิดล โครงการวิจัย วัณโรค ระดับนานาชาติ พัฒนา เทคโนโลยี นวัตกรรม ทางวิทยาศาสตร์การแพทย์ สำหรับ ตรวจหา ลักษณะทางพันธุกรรม มนุษย์ เชื้อ วัณโรค วินิจฉัย วัณโรค แม่นยำ เลือก ยา ขนาดยา ต้าน วัณโรค เหมาะสม ลด อาการ พึงประสงค์ ยา ต้าน วัณโรค นโยบาย ยุติ วัณโรค	ข่าวจริง
⋮	⋮	⋮	⋮
3195	หมอเตือน "ไข้หวัดใหญ่" ระบาดหนักแน่ หลังป่วยสูงตั้งแต่ต้นปีรวมกว่า 1.52 แสนราย ตาย 10 ราย คาดป่วย 2 แสนรายเป็นอย่างต่ำ แต่ตัวเลขจริงอาจถึงล้านคน ห่วงฤดูฝน-เปิดเทอมยิ่งเพิ่มการระบาด	หมอ เตือน ไข้หวัดใหญ่ ระบาด หนัก แน่ ป่วย ต้นปี ห้า สิบสอง แสน ตาย สิบ คาด ป่วย สอง แสน ต่ำ ตัวเลข ล้าน คน ห่วง ฤดูฝน เปิดเทอม ระบาด	ข่าวจริง
3196	ในช่วงการระบาดของโควิด-19 การรักษาสุขอนามัยถือเป็นสิ่งสำคัญ โดยเฉพาะอาหารและน้ำ ที่ต้องสะอาดและปลอดภัยต่อการบริโภค ขณะที่ผู้ประกอบการและผู้สัมผัสอาหาร ควรเรียนรู้หลักสุขาภิบาลอาหาร	ระบาด โควิด สิบ เก้า การรักษา สุขอนามัย ถือเป็น โดยเฉพาะ อาหาร น้ำ สะอาด ปลอดภัย บริโภค ผู้ประกอบการ สัมผัส อาหาร เรียนรู้ หลัก สุขาภิบาล อาหาร	ข่าวจริง

Table 17 Text Preprocessing Results for Case 8

No.	Topic	Topic_After	Label
3197	กรมอนามัย ส่งเสริมหญิงตั้งครรภ์ใช้สมุดบันทึกสุขภาพแม่และเด็ก เล่มสีชมพู เป็นเครื่องมือในการดูแลสุขภาพตนเองและลูกในครรภ์ ช่วงตั้งครรภ์อย่างต่อเนื่องไปจนเด็กเข้าโรงเรียน	กรมอนามัย ส่งเสริม หญิง ตั้งครรภ์ สมุดบันทึก สุขภาพ แม่ เด็ก เล่ม สีชมพู เครื่องมือ ดูแล สุขภาพ ลูกครรภ์ ตั้งครรภ์ ต่อเนื่อง เด็ก โรงเรียน	ข่าวจริง
3198	ซื้อ-ขายใบกระท่อม ไม่ต้องขออนุญาตกับทางอย.	ซื้อขาย ใบ กระท่อม ไม่ต้อง ขออนุญาต ทา งอย	ข่าวจริง
3199	หากรพ.ขอรับบริจาคอุปกรณ์ทางการแพทย์ให้แจ้งมาที่องค์การเภสัชกรรม	รพ รับบริจาค อุปกรณ์ ทางการแพทย์ แจ้ง องค์การเภสัชกรรม	ข่าวปลอม

4.2 Accuracy of the Three Models for Different Text Preprocessing Cases

The Naïve Bayes model achieved the highest accuracy of 0.9344 in case 2, which only involved replacing numbers with Thai words. The lowest accuracy of 0.9156 occurred in cases 6-8, all of which included removing stop-words. Removing punctuation had minimal impact on accuracy. These results suggest that for the Naïve Bayes model, removing numbers improved performance slightly, while removing stop-words decreased accuracy. Punctuation removal was largely inconsequential, as shown in Figure 5.

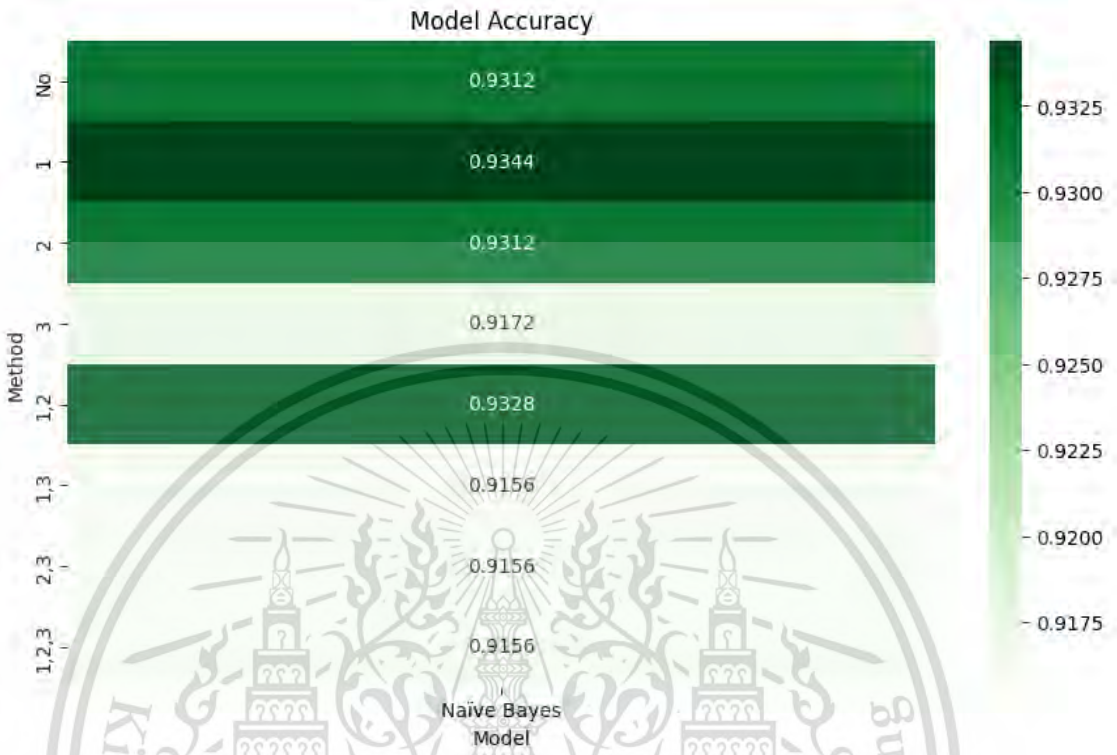

Figure 5 Naïve Bayes Model Accuracy for Different Text Preprocessing Cases

The SVM model showed the best accuracy of 0.9547 in case 2, again using only replacing numbers with Thai words. The lowest accuracy of 0.9266 was in case 4, which only removing stop-words. Like with Naïve Bayes, removing numbers boosted accuracy while removing stop-words reduced it, and punctuation removal had little effect. However, the accuracy changes were more pronounced for SVM compared to Naïve Bayes, as shown in Figure 6.

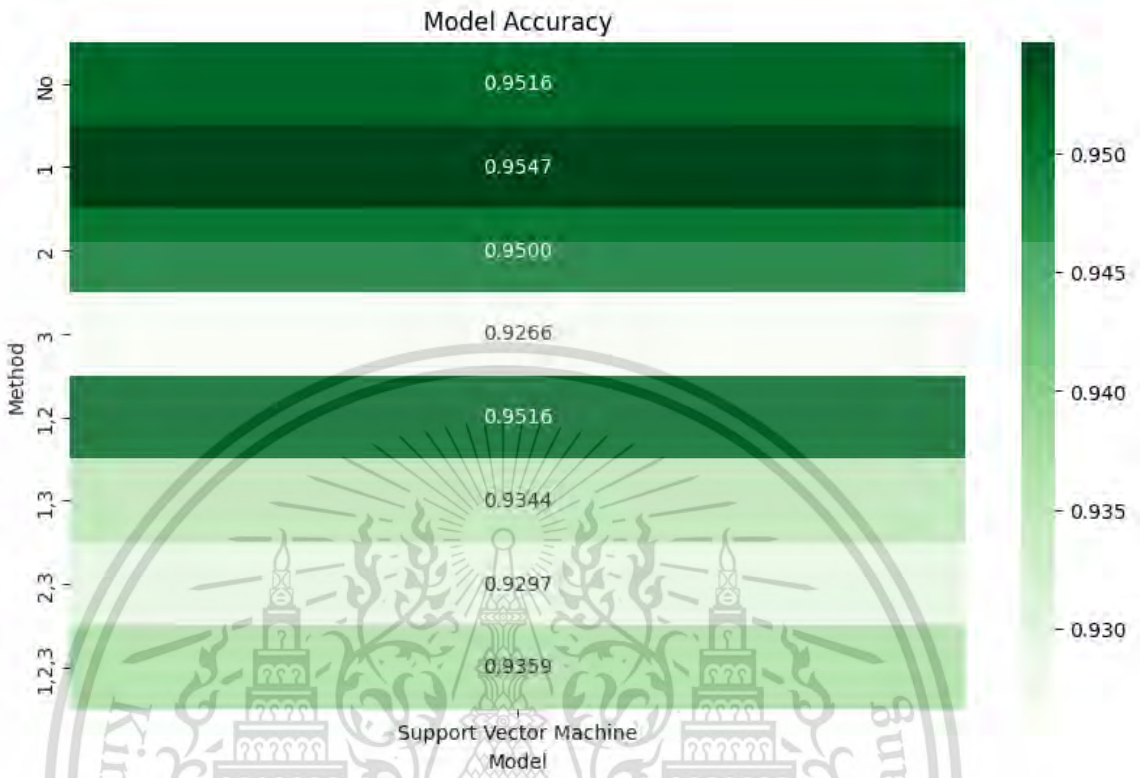

Figure 6 SVM Model Accuracy for Different Text Preprocessing Cases

Random Forest achieved peak accuracies of 0.9297 in both case 2 (Replacing numbers with Thai words only) and case 5 (Replacing numbers with Thai words and Removing punctuation). The worst accuracy of 0.9047 was in case 6 (Replacing numbers with Thai words and Removing stop-words). Once again, number removal improved accuracy, stop-word removal decreased it, and punctuation removal was mostly neutral. The accuracy fluctuations were smaller than for SVM but larger than Naïve Bayes, as shown in Figure 7.

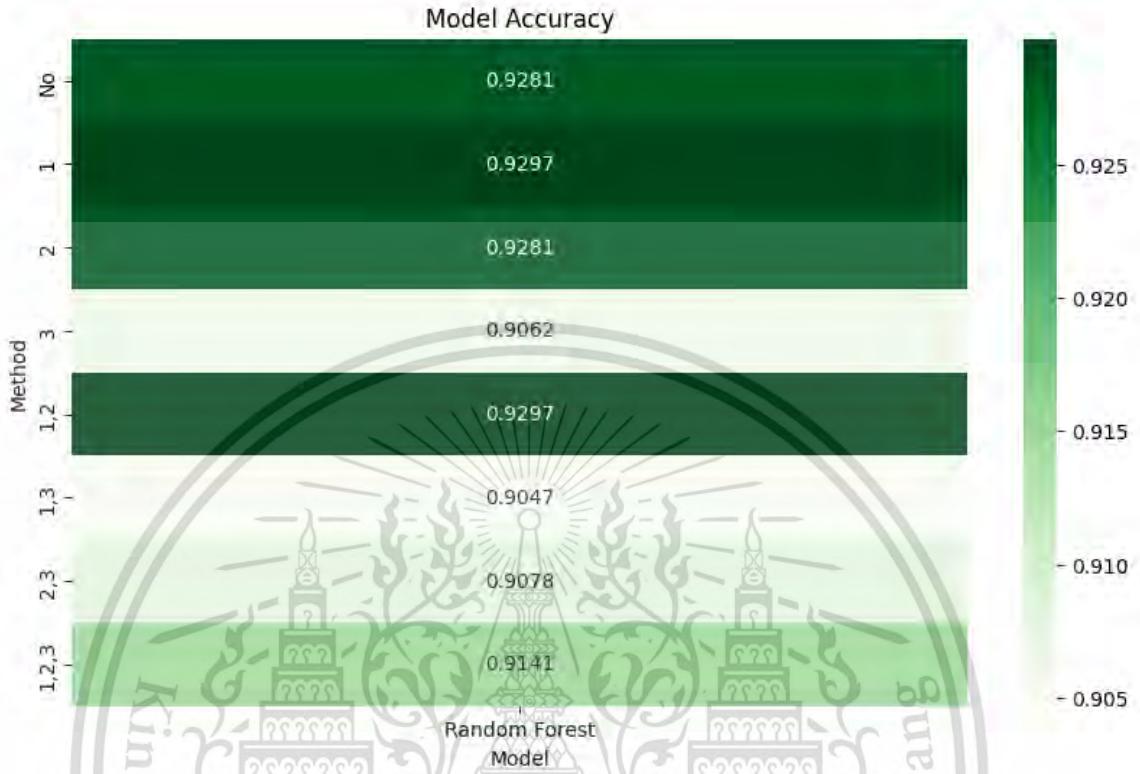

Figure 7 Random Forest Model Accuracy for Different Text Preprocessing Cases

Across all three models, the best results occurred when using only text preprocessing method 1 (replacing numbers with Thai words), with peak accuracies of 0.9344 for Naïve Bayes, 0.9547 for SVM, and 0.9297 for Random Forest. The lowest accuracies for each model were seen when stop-word removal was included.

The preprocessing techniques applied to the Thai health news dataset revealed several key insights. First, removing numbers that were accompanied by Thai words led to consistent improvements in accuracy across all models tested, suggesting that the presence of numbers in this context may introduce noise and hinder classification performance. Second, contrary to expectations, removing stop-words resulted in decreased accuracy for all models, indicating that these commonly occurring words contain valuable information for distinguishing between real and fake health news in the Thai language. Third, the removal of punctuation had minimal

impact on the models' performance, implying that punctuation does not play a significant role in differentiating between the two classes of news articles. Finally, the Support Vector Machine (SVM) model demonstrated the highest overall accuracy and was the most sensitive to changes in preprocessing techniques, followed by the Random Forest model and then the Naïve Bayes model.

Table 18 Accuracy of the Three Models for Different Text Preprocessing Cases

Case	Method	Naïve Bayes	Support Vector Machine	Random Forest
1	No	0.9312	0.9516	0.9281
2	1	0.9344	0.9547	0.9297
3	2	0.9312	0.9500	0.9281
4	3	0.9172	0.9266	0.9062
5	1,2	0.9328	0.9516	0.9297
6	1,3	0.9156	0.9344	0.9047
7	2,3	0.9156	0.9297	0.9078
8	1,2,3	0.9156	0.9359	0.9141

Remark: 1 = Replacing Numbers with Thai Words, 2 = Removing punctuation, 3 = Removing stop-words

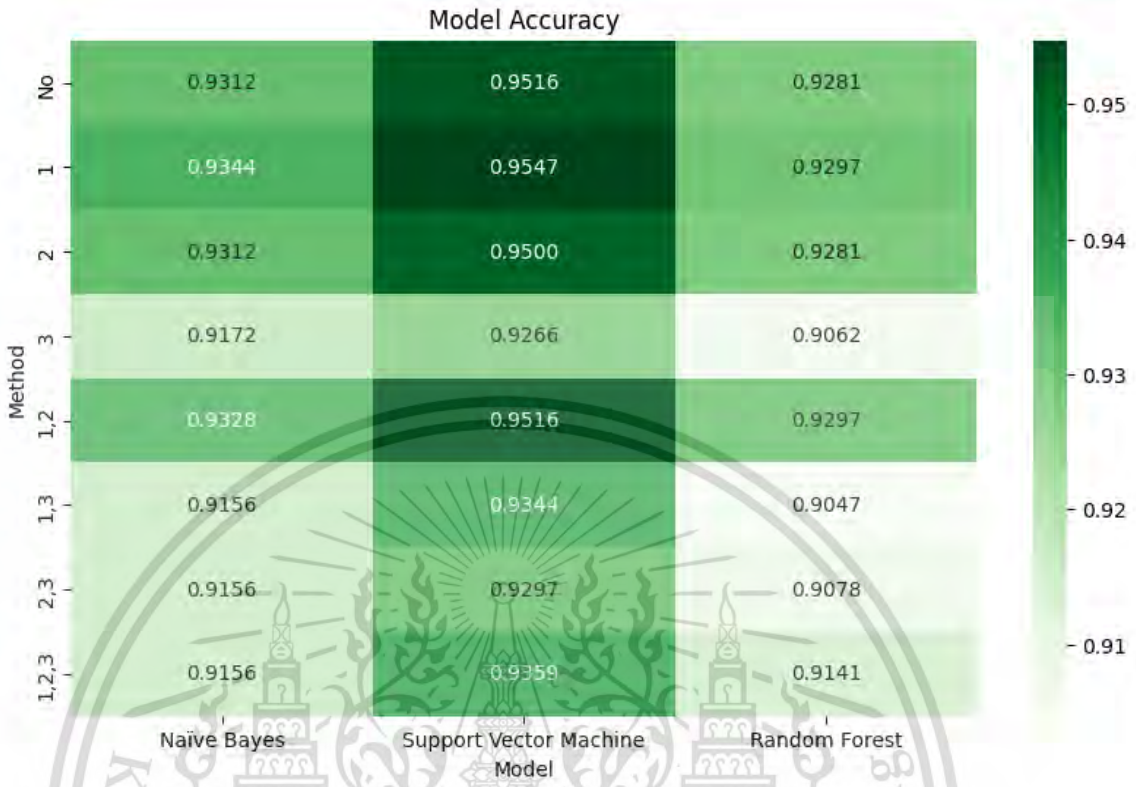

Figure 8 Accuracy of the Three Models for Different Text Preprocessing Cases

Chapter 5
Conclusion and Suggestion

Chapter 5 is divided into two parts: an analysis of the results from Chapter 4 and future research directions.

5.1 Conclusion

When examining the impact of changing text preprocessing methods on model accuracy, it was found that for all models, the Replacing Numbers with Thai Words method yielded the best results and outperformed the case where none of the three text preprocessing methods were used. This is likely due to the relatively high accuracy of converting numbers to Thai text using PyThaiNLP.

This was evident even in cases involving years such as "2562" being correctly converted to "สอง พัน ห้า ร้อย หก สิบสอง". However, the Replacing Numbers with Thai Words method providing the best results, there were still minor differences. This may be because, in some cases, the conversion is not entirely accurate, such as converting "1.52" to "หนึ่ง.ห้าสิบสอง" instead of "หนึ่ง.ห้าสอง".

The Removing Punctuation method yielded similar results to the case where none of the three text preprocessing methods were used. The reason why this method did not improve accuracy is likely because removing punctuation may cause some parts of the sentence to lose meaning. For example, removing the "ไม้ยมก" symbol, which is used to indicate word repetition, may reduce the occurrence of repeated words.

Lastly, the Removing Stop Words method resulted in the lowest accuracy compared to the other two methods and even performed worse than the case where none of the three text preprocessing methods were used. This is likely because the list of stop-words used for word removal was not specifically designed for misinformation detection but was taken from PyThaiNLP. As a result, in some cases, words with significant meaning may have been removed, such as "ยิ่งใหญ่" or "นิดหน่อย".

When comparing the models used, it was found that when using the same text preprocessing methods, Support Vector Machine provided the highest accuracy compared to Naïve Bayes and Random Forest. This is likely because SVMs generally perform well with smaller datasets compared to other algorithms and are effective when there is a clear separation between classes in the feature space.

5.2 Limitation

The experiments in this research were conducted on a dataset with a limited size, which may affect the generalizability of the findings. The study focused on Thai health-related fake news, and the effectiveness of the text preprocessing methods may vary when applied to fake news in other categories or languages.

5.3 Suggestion

The suggestions are divided into two parts: recommendations for improving the three text preprocessing methods used in this research and other interesting suggestions for future experimentation.

5.3.1 Suggestion for Enhancing the Text Preprocessing Methods Used in this Research

In the future, the three text preprocessing methods selected for this research can be further explored. For the Replacing Numbers with Thai Words method, experiments can be conducted to find ways to convert numbers to Thai words while considering the context of Thai pronunciation. For example, converting "1.52" to "หนึ่ง. ห้าสอง". For the Removing Punctuation method, experiments can be conducted by converting punctuation marks that affect meaning into Thai words. For instance, converting the "ไม้ยมก" symbol to repeated word or converting the "มหัพภาค" symbol to the word "จุด". Lastly, the Removing Stop Words method can be further explored by creating a list of words that are suitable for misinformation detection.

5.3.2 Expanding the Scope of Experimentation

In the future, additional experiments can be performed using a larger dataset of Thai health-related fake news, as well as fake Thai news in other categories. Moreover, the impact of text preprocessing methods other than the three used in this research can be explored, such as converting commonly used abbreviations to their full names, translating English words to Thai, and correcting misspelled words.

REFERENCES

[1] Anti - Fake News Center. **รายงานผลการดำเนินงาน AFNC ประจำปี 2566.** [Online]. Available: https://www.mdes.go.th/storage/contents/file/QywyWbTF6t3OBMwqaDp7dgszCnm9P75dkU8QT8Cf.pdf

[2] El Mikati IK, Hoteit R, Harb T, El Zein O, Piggott T, Melki J, Mustafa RA and Akl EA. (2023). **Defining Misinformation and Related Terms in Health-Related Literature: Scoping Review**. doi: https://www.jmir.org/2023/1/e45731/.2023.

[3] Cambridge Dictionary. **Meaning of fake news.** [Online]. Available: https://dictionary.cambridge.org/dictionary/english/fake-news

[4] Dictionary.com. **Meaning of misinformation.** [Online]. Available: https://www.dictionary.com/browse/misinformation

[5] Dictionary.com. **Meaning of disinformation** [Online]. Available: https://www.dictionary.com/browse/disinformation

[6] National Library of Australia. **What is fake news, misinformation, and disinformation?** [Online]. Available: https://www.nla.gov.au/faq/what-is-fake-news
-misinformation-and-disinformation

[7] University of Washington Bothell & Cascadia College Campus Library. **News: Fake News, Misinformation & Disinformation.** [Online]. Available : https://guides.lib.uw.edu/bothell/news/misinfo

[8] Sunil Ray. **Naive Bayes Classifier Explained: Applications and Practice Problems of Naive Bayes Classifier.** [Online]. Available: https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/

[9] Rohith Gandhi. **Naive Bayes Classifier.** [Online]. Available: https://towardsdatascience.com/naive-bayes-classifier-81d512f50a7c

[10] IBM. **What are Naïve Bayes classifiers?.** [Online]. Available: https://www.ibm.com/topics/naive-bayes

[11] Sunil Ray. **Learn How to Use Support Vector Machines (SVM) for Data Science.** [Online]. Available: https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/

[12] Rohith Gandhi. **Support Vector Machine — Introduction to Machine Learning Algorithms.** [Online]. Available : https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47

[13] IBM. **What are support vector machines (SVMs)?**. [Online]. Available: https://www.ibm.com/topics/support-vector-machine

[14] Sruthi E R. **Understand Random Forest Algorithms with Examples (Updated 2024).** [Online]. Available: https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/

[15] Niklas Donges. **Random Forest: A Complete Guide for Machine Learning.** [Online]. Available: https://builtin.com/data-science/random-forest-algorithm

[16] IBM. **What is random forest?.** [Online]. Available: https://www.ibm.com/topics/random-forest

[17] Aniruddha Bhandari. **Understanding & Interpreting Confusion Matrix in Machine Learning (Updated 2024).** [Online]. Available : https://www.analyticsvidhya.com/blog/2020/04/confusion-matrix-machine-learning/

[18] GeeksforGeeks. **Confusion Matrix in Machine Learning.** [Online]. Available : https://www.geeksforgeeks.org/confusion-matrix-machine-learning/

[19] Nisha Arya Ahmed. **What is A Confusion Matrix in Machine Learning? The Model Evaluation Tool Explained.** [Online]. Available: https://www.datacamp.com/tutorial/what-is-a-confusion-matrix-in-machine-learning

[20] Aysel Aydin. **1 — Text Preprocessing Techniques for NLP.** [Online]. Available: https://ayselaydin.medium.com/1-text-preprocessing-techniques-for-nlp-37544483c007

[21] Aysel Aydin. **2— Stemming & Lemmatization in NLP: Text Preprocessing Techniques.** [Online]. Available : https://ayselaydin.medium.com/2-stemming-lemmatization-in-nlp-text-preprocessing-techniques-adfe4d84ceee

[22] Maleesha De Silva. **Preprocessing Steps for Natural Language Processing (NLP): A Beginner's Guide.** [Online]. Available: https://medium.com/@maleeshadesilva21/preprocessing-steps-for-natural-language-processing-nlp-a-beginners-guide-d6d9bf7689c9

[23] Tithi Sreemany. **Essential Text Pre-processing Techniques for NLP!.** [Online]. Available: https://www.analyticsvidhya.com/blog/2021/09/essential-text-pre-processing-techniques-for-nlp/

[24] Pranshav Patel. **Unlocking the Power of NLP: A Deep Dive into Text Preprocessing Steps.** [Online]. Available: https://medium.com/@pranshavpatel/unlocking-the-power-of-nlp-a-deep-dive-into-text-preprocessing-steps-8eb5dfe8b94

[25] Supanya Aphiwongsophon. **DETECTING FAKE NEWS WITH MACHINE LEARNING METHOD** doi: https://digital.car.chula.ac.th/chulaetd/3397/

[26] Pakpoom Mookdarsanit and Lawankorn Mookdarsanit. **The COVID-19 fake news detection in Thai social texts.** doi: https://www.researchgate.net/publication/350560656_The_COVID-19_fake_news_detection_in_Thai_social_texts

[27] Phayung Meesad, Phnom Kleechaya, Aongart Aun-a-nan and Kamonrat Kijrungpaisarn. **Artificial Intelligent Techniques for Thai Fake News Detection.** doi: https://ph01.tci-thaijo.org/index.php/JASCI/article/view/244590

[28] Rutchaneewan Kowirat. **FAKE NEWS Detection on social media: case study of coronavirus 2019.** doi: https://dl.acm.org/doi/10.1145/3510249.3510301

[29] Kotchakorn Tiemtud. **THAI FAKE NEWS DETECTION USING MACHINE LEARNING MODEL.** [Online]. doi: https://digital.library.tu.ac.th/tu_dc/frontend/Info/item/dc:305738

[30] Akrivi Krouska, Christos Troussas and Maria Virvou. **The effect of preprocessing techniques on Twitter Sentiment Analysis.** doi: https://www.researchgate.net/publication/311755864_The_effect_of_preprocessing_techniques_on_Twitter_sentiment_analysis

[31] Yaakov HaCohen-KernerID and Daniel MillerID, Yair Yigal. **The influence of preprocessing on text classification using a bag-of-words representation.** doi: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0232525

[32] Wilson Fearn, Orion Weller, Kevin Seppi. **Exploring the Relationship Between Algorithm Performance, Vocabulary, and Run-Time in Text Classification.** doi: https://arxiv.org/abs/2104.03848

[33] Witchapong Daroontham. **ขั้นตอนการเตรียมข้อมูลประเภท Text ภาษาไทย แบบง่ายๆ โดยใช้ Python (Simple Thai text preprocessing using Python).** [Online]. Available: https://medium.com/@witchapongdaroontham/ขั้นตอนการเตรียมข้อมูลประเภท-text-ภาษาไทย-แบบง่ายๆ-โดยใช้-python-simple-thai-text-preprocessing-c8c46ca3ce46

[34] PyThaiNLP. **"What is PyThaiNLP?"** [Online]. Available: https://pythainlp.github.io/FAQ

AUTHOR BIOGRAPHY

Name	Weeranuch Proysaithong
Email Address	weeramuch.pt@gmail.com
Educational Background	Bachelor of Economics, Thammasat University, GPA: 3.26