Journal of Cases on Information Technology
Volume 26 • Issue 1

Improvement of a Machine Learning Model Using a Sentiment Analysis Algorithm to Detect Fake News:

A Case Study of Health and Medical Articles on Thai Language Websites

Kanokwan Atchariyachanvanich, School of Information Technology, King Mongkut's Institute of Technology Ladkrabang, Thailand*
https://orcid.org/0000-0002-2705-7942

Chotipong Saengkhunthod, School of Information Technology, King Mongkut's Institute of Technology Ladkrabang, Thailand

Parischaya Kerdnoonwong, School of Information Technology, King Mongkut's Institute of Technology Ladkrabang, Thailand

Hutchatai Chanlekha, Department of Computer Engineering, Faculty of Engineering, Kasetsart University, Thailand
https://orcid.org/0009-0005-2607-1936

Nagul Cooharojananone, Department of Mathematics and Computer Science, Faculty of Science, Chulalongkorn University, Thailand

ABSTRACT

These days, the problem of fake news has grown to be a major social and personal concern. With the amount of information generated through social media, it is very crucial to be able to detect and properly take care of that fake information. Previous studies proposed a machine learning model to detect fake news in online Thai health and medical articles. Still, the problem of detecting fake news with similar content but different objectives exists, and the accuracy of the model needs improvement. Therefore, this study aims to solve these problems by adding 33 new features, including textual features, sentiment-based features, and lexicon features, i.e., herbs, fruits, and vegetables, to identify the objective of an article. We trained and tested the model's prediction accuracy on a new dataset containing 582 reliable and 435 unreliable (fake news) articles from eight Thai websites. Our improved classification model using XGBoost with Lasso, the best feature selection method, achieved an accuracy of 97.76% without over-fitting, reflecting a 7.16% improvement over our earlier model.

KEYWORDS

Data Science, Health Informatics, Lexicon Features, Misinformation, Thai-Text Analysis

DOI: 10.4018/JCIT.344812

*Corresponding Author

The invention of social media has made it easier for people to disseminate and spread news, which has sped up and expanded the circulation of many sorts of content. In Thailand, fake news has not been officially defined. The spread of fake news has far-reaching implications that can cause misunderstanding among people. The Electronic Transactions Development Agency (ETDA)'s 2020 survey of Internet behavior in Thailand found that Thai people now spent an average of 11 hours and 25 minutes per day on the Internet in 2020, up one hour and three minutes from 2019 (Electronic Transactions Development Agency (ETDA), 2020). Thai Internet users actively consume and share news from online websites or social media. When consuming news on social media, they may be unaware that they are receiving information from news sites that were created to spread false ideas. About 70% of ETDA respondents found fake news about health or health products on social media (Electronic Transactions Development Agency (ETDA), 2020). Fake news can mislead and influence people to adopt behaviors that can lead to health problems. According to the Anti-Fake News Center of Thailand, from November 2019 to October 2021, 53% of fake news articles involved health information, the highest among all categories (Ministry of Digital Economy and Society (MDES), n.d.a). Based on a guideline for tracking fake news information, the Anti-Fake News Center of Thailand focuses on news and information that directly affects the lives and properties of Thai people (MDES, n.d.a). Therefore, this research focuses on health-related fake news, which is defined as any factually inaccurate health or medical article that directly affects the lives and health of people. Fake news is cheap to provide online and is quickly disseminated by readers through social media, usually without filtering or first verifying its accuracy (Shu et al., 2017), while accessing information on the Internet is an essential part of modern human life.

The researchers observed the characteristics of Thai health and medical articles. Most reliable articles are written by a trusted healthcare professional or by a medical institution and are written to clear up any misunderstanding from other articles. The latter characteristic aims to refute the fake news articles. These kind of reliable articles usually contain contents paraphrased from fake news articles together with contents generated from reliable sources. This paraphrased text may cause a reliable article to be classified as fake news articles. This causes a problem of misclassifying health and medical articles that have similar content but serve different purposes. Moreover, the researchers observed that articles refuting fake news or articles from reliable sources often contain negative sentiments such as: "ไม่จริง (untrue)," "อย่าแชร์ (do not share)," and "ไม่เชื่อ (do not believe)"; whereas unreliable articles contained positive messages such as "´ดีจริง ๆ (really good)" and ของดี (good stuff)." Moreover, many unreliable health and medical articles usually contain information about fruit, vegetables, herbs, diseases, and body organs. These words are also found in many original fake news articles posted on the Anti-Fake News Center of Thailand website (MDES, n.d.a).

Machine learning (ML) algorithms have been applied to classify fake news (Aphiwongsophon & Chongstitvatana, 2018; Aslam et al., 2021; Dey et al., 2018; Mookdarsanit & Mookdarsanit, 2021; Nyow & Chua, 2019; Ozbay & Alatas, 2020; Reis et al., 2019), or reliable/unreliable information in social media (Liu et al., 2019; Saengkhunthod, Kerdnoonwong, & Atchariyachanvanich, 2021). The researchers' previous model to predict unreliable medical articles on Thai websites presented 20 features that affected the reliability or unreliability of the articles, and XGBoost methods were the most effective at 90.60% accuracy (Saengkhunthod et al., 2021). However, the result from the previous model (Saengkhunthod et al., 2021) was unable to classify health and medical reliable articles that refuted the fake news articles. Research has been done on sentiment features for fake news detection, such as percent of positive words and percent of negative words (Ajao, Bhowmik, & Zargari, 2019; Zhou & Zafarani, 2020). Words can be identified as either good or bad at displaying either positive or negative emotions by looking at the keywords used in the texts posted online (Ajao et al., 2019). There is an improvement in detecting rumor in a Tweet when adding an emotional word feature in the classifier (Ajao et al., 2019). The best results for sentiment-aware text-only rumor detection was achieved at 86% accuracy by the support vector machines (SVMs) and the hierarchical attention networks (HAN) models, with Twitter pre-trained word embedding (Ajao et al., 2019). The sentiment analysis was also used to extract polarity and subjectivity from given Tweets to locate the deceptive words in the domain of political news (Dey et al., 2018). Although unreliable, Thai health and medical articles usually contain sentiment words, to the researchers' knowledge, there has been no research that utilized such a sentiment feature in identifying medical-related fake news. Therefore, this sentiment analysis of emotional words in the texts can be applied in the context of Thai health and medical articles to help classifying reliable articles having negative sentiments fake news with positive sentiments in it, and positive sentiments with similar content, but with different objectives, since the reliable articles often contain negative sentiments, whereas unreliable articles contained positive messages.

Here, the researchers' objective was to improve the efficiency of their previous model (Saengkhunthod et al., 2021) by collecting additional datasets and adding new domain-specific lexicon features, such as herbs, vegetables, disease, etc. A sentiment analysis technique was first applied to solve the problem of detecting fake news in the domain of health and medical articles from Thai websites. The researchers' new model focuses on classifying articles as unreliable (fake news) or reliable.

The following are the three main contributions of this paper:

- The researchers propose new features on herbs, vegetables, diseases, and body organs found in health and medical articles on Thai language websites to help classifying the fake news.
- The researchers originally apply a sentiment analysis to solve the problem of detecting fake news with similar content, but with different objectives in health and medical articles on Thai language websites.
- The researchers developed the ML model using new features and a sentiment-based feature that is indicated to improve the performance of the previous model.

LITERATURE REVIEW

Approaches to Identify Fake News in Health and Medical Articles

While many definitions encompass the broader phenomenon of fake news (Fake-news, n.d.a; Fake-news, n.d.b; Molina et al., 2021; Wu et al., 2019), various checks can be used to detect fake news in health and medical reporting (Treharne & Papanikitas, 2020). There are four areas of validation that can help identify fake news stories. First, the news story should be searched on the media publication's official site or, if available, in the hardcopy newspaper to verify its authenticity from the original source. Second, the reader should check to see if the content in question appears on other reputable websites. This is referred to as "scope of coverage." Third, fact-checking sites, such as Snopes.com and Factcheck.org, which list current fake news stories, should be consulted. Finally, a generic search of the publication title should be conducted to see if the news item is from a parody publication (Treharne & Papanikitas, 2020).

In Thailand, the Antifakenewscenter.com website is maintained by the Anti-Fake News Center managed by the Ministry of Digital Economy and Society (MDES, n.d.a). This agency collects news on various topics, verifies their content as fake news (or not), and then publishes them on the website. The agency aims to help people become aware of fake news and helps prevent the spread of fake news (MDES, n.d.a). The center labels the types of fake news that have a wide impact, because fake news directly affects people's lives and assets, creates social divisions and misconceptions about society, and destroys the image of the country (Shu et al., 2017).

Figure 1 illustrates an example of a health and medical fake news article from the Anti-Fake News Thailand website. The title, "Lime Soda Cures Cancer," went viral on social media. Articles in Thailand are popular for educating people on cures for various diseases. If the published articles are unreliable, they may harm people reading them because an unsuitable diet has a negative impact

Figure 1. Typical Fake News Article: The Title Translates As "Lime Soda Cures Cancer," Taken From the Government Anti-Fake News Center (MDES, n.d.b)

ข่าวปลอม อย่าแชร์! มะนาวโซดารักษาโรคมะเร็ง

18 สิงหาคม 2020 | 17:01

Share Facebook Twitter Email Link Youtube Tiktok

ตามที่มีข้อความแนะนำ เกี่ยวกับประเด็นเรื่อง **มะนาวโซดารักษาโรคมะเร็ง** ทางศูนย์ต่อต้านข่าวปลอมได้ดำเนินการตรวจสอบข้อเท็จจริงโดย สำนักงานคณะกรรมการอาหารและยา กระทรวงสาธารณสุข พบว่าประเด็นดังกล่าวนั้น **เป็นข้อมูลเท็จ**

จากที่มีข้อความแนะนำว่าให้ดื่มน้ำมะนาวผสมโซดา เพื่อช่วยฆ่าเซลล์มะเร็งนั้น ทางสำนักงานคณะกรรมการอาหารและยาได้ชี้แจงว่า **น้ำมะนาว และน้ำโซดา ไม่มีข้อบ่งใช้ในการรักษาโรคมะเร็ง** โดยน้ำมะนาว มีส่วนประกอบหลักคือ กรดซิตริก (Citric acid) เป็นกรดอย่างอ่อน มีรสเปรี้ยว จะกระตุ้นให้มีการขับน้ำลายออกมาทำให้ชุ่มคอ ช่วยบรรเทาอาการเจ็บคอ ส่วนน้ำโซดา (Carbonate beverages) เกิดจากการนำเอาก๊าซคาร์บอนไดออกไซด์ มาอัดลงในน้ำ ทำให้ส่วนประกอบหลักของน้ำโซดาคือ กรดคาร์บอเนต มีความเป็นกรดเช่นเดียวกับกรดซิตริก ซึ่งมีบางงานวิจัย กล่าวว่า การดื่มน้ำโซดาในขณะท้องว่าง หรือการดื่มน้ำโซดาในปริมาณที่มากเกินไป จะทำให้เกิดผลเสียต่าง ๆ ต่อระบบทางเดินอาหารตามมา ไม่ว่าจะเป็นการไม่สบายท้อง เกิดภาวะกรดไหลย้อน หรือเกิดการระคายเคืองเยื่อบุในระบบทางเดินอาหาร

on the body. Vegetables, fruits, and herbs are frequently used to treat various diseases. In addition, unreliable articles often refer to specific diseases such as cancer, diseases of the brain, and other organs.

Characteristics of Health and Medical Websites

Samuel and Zaiane (2012) reported that most Internet users tend to believe articles in public online media without verification, and often use medical advice from websites. They evaluated the credibility of article-based websites written by an expert authority, and community-based websites using privacy, security, and trust as the primary indicators/constructs. Information provided by article-based websites is written by experts, therefore each article can be validated by cross checking with a confirmed valid article (Samuel & Zaiane, 2012). Molina et al. (2021) stated that a reliable news feature should be fact-checked, impartial, and should have clear citations with accurate statistical data derived from a relevant research organization or authority. In contrast, fake news presents incomplete information, the writing quality is often poor, spelling errors are common, and it may include images that are unrelated to the topic of the article. Finally, the sources of the information presented are either not provided, or they are incorrect (Molina et al., 2021). Zhou and Zafarani (2020) suggested that the reliability of news sources, i.e., author, publisher, and user, who spread the news stories on social media, should be assessed to detect fake news.

The researchers found that many reliable articles had been written to correct misunderstandings in unreliable articles on the same subject of health and medical. For example, two articles discussed the same fruit, but one article aimed to promote the health benefits of that fruit, and another aimed to counter those claims. These two types of health and medical articles have a similar focus (e.g., fruit, disease, and symptoms), but their objectives were different. Consequently, the sentiment analysis technique was applied to classify similar articles with different objectives.

Previous Works on Automatic Fake News Detection

Since fake news is created with the goal of misleading readers and imitating reliable news sources, it can be difficult to tell the difference between it and actual news (Singh et al., 2023). Research on fake news detection has been addressed since 2011 (Singh et al., 2023). Lui et al. (2019) analyzed the context and nature of reliable and unreliable news on Chinese social media by exploring the writing style, topics, and numbers of special characters to identify differences in these types of articles. However, fake news in other domains such as politics (Dey et al., 2018; Mookdarsanit & Mookdarsanit, 2021; Ozbay & Alatas, 2020; Reis et al., 2019), and natural phenomena (Aphiwongsophon & Chongstitvatana, 2018) were detected and analyzed using several methods. ML methods for modeling use classification models, including SVMs, k-nearest neighbors (kNN), naïve Bayes, AdaBoost, gradient-boosted decision trees (GBDT), and Random Forests (RFs) (Aslam et al., 2021; Dey et al., 2018; Nyow & Chua, 2019; Ozbay & Alatas, 2020; Reis et al., 2019), and detected fake news on Twitter using machine-learning models and showed that RFs achieved the best results, with an F1 score of 97.2% (Nyow & Chua, 2019). Deep learning models, including the bidirectional long short-term memory and gated recurrent unit with dense layers (Bi-LSTM-GRU-dense), bidirectional encoder representations from transformers (BERT), universal language model fine-tuning (ULMFiT), and generative pre-trained transformers (GPT) (Aslam et al., 2021), have also been used to analyze the data. The deep learning model with Bi-LSTM-GRU-dense was the most effective, with an 89% accuracy, whereas the ULMFiT model achieved only 72%. Still, these models contain a large amount of noise and generate many outliers, so that the supervised learning model was unable to predict accurately. A unique, two-channel deep-learning framework named HANCaps, for identifying fake news in Thai, was implemented by capturing the hierarchical relationships encoded within textual features, and making use of the combination of the HAN and capsule networks with BERT and FastText embeddings (Maity et al., 2023).

Nadeem et al. (2023) presented HyproBert, a hybrid model for fake news detection based on deep hypercontext. It leverages the distill-Bert for embeddings, the convolution neural network for extracting spatial data, the bidirectional gated recurrent unit (BiGRU) for extracting contextual data, and the self-attention-capable CapsNet for hierarchical comprehension of both complete and partial relations among data. Palani and Elango (2023) proposed the content-based ensemble of a deep learning-based framework, named the BERT-BiLSTM-convolutional neural network (CNN) for Fake News Detection (BBC-FND). Its performance was evaluated using four benchmark datasets for fake news: McIntire, Covid-19, Kaggle, and WELFake. The findings demonstrated that the BBC-FND model performed better than the other state-of-the-art (SoTA) methods, with accuracy on four datasets of 97.31%, 98.64%, 99.06%, and 98.26%, respectively. Mookdarsanit and Mookdarsanit (2021) recommended that fake news detection be semi-supervised, and that the dataset should be trained using a partially supervised model to select only high-quality labels in Thai text and label unknown Thai texts to decrease the model training time (Mookdarsanit & Mookdarsanit, 2021). Aphiwongsophon and Chongstitvatana (2018) used neural networks to detect fake news in Thailand and reported a 99.9% accuracy. Songram (2019) and colleagues selected features using the Gini index, χ^2, Fisher and least absolute shrinkage and selection operator (LASSO) to explore the characteristics of unreliable and reliable news pages on Facebook.

Sentiment analysis was applied with the proposed sentiment features, linguistic features, and named entity-based features to be trained in deep learning models such as gated recurrent unit (GRU), long short-term memory (LSTM), and recurrent neural network (RNN) for detecting fake news about Covid-19 (Iwendi et al., 2022). Jadhav and Shukla (2024) introduced a deep learning analysis using linguistic complexity and semantic signature to identify fake news. The BERT exhibits remarkable precision and nuanced semantics, whereas the LSTM-Attention Mechanism is excellent at preserving semantic consistency. Furthermore, GPT provides a large selection of word representations (Jadhav & Shukla, 2024). In order to identify false information from web URLs, Barve et al. (2022) presented a novel sentiment-based incremental machine-learning approach. Zhou and Zafarani (2020) found that sentiment was one of the latent textual features in machine-learning models for the detection of fake news. Ajao et al. (2019) used sentiment analysis to detect fake news on social media. They hypothesized a relationship between sentiments, false statements, and rumors, and found that fake news writers often use negative messages that spread quickly. Therefore, they used sentiment analysis to calculate an emotional score using latent semantic analysis (LSA) and latent Dirichlet allocation (LDA). ML was used for prediction, and SVM models were the most effective, with an 86% accuracy. Barve and Saini (2023) proposed a technique called "Content Similarity Measure (CSM)" that may automatically verify the accuracy of URLs in the healthcare industry. To accomplish journalistic fact checking, authors have added a unique collection of sentiment polarity, domain-specific, and content similarity score features.

Baseline Model

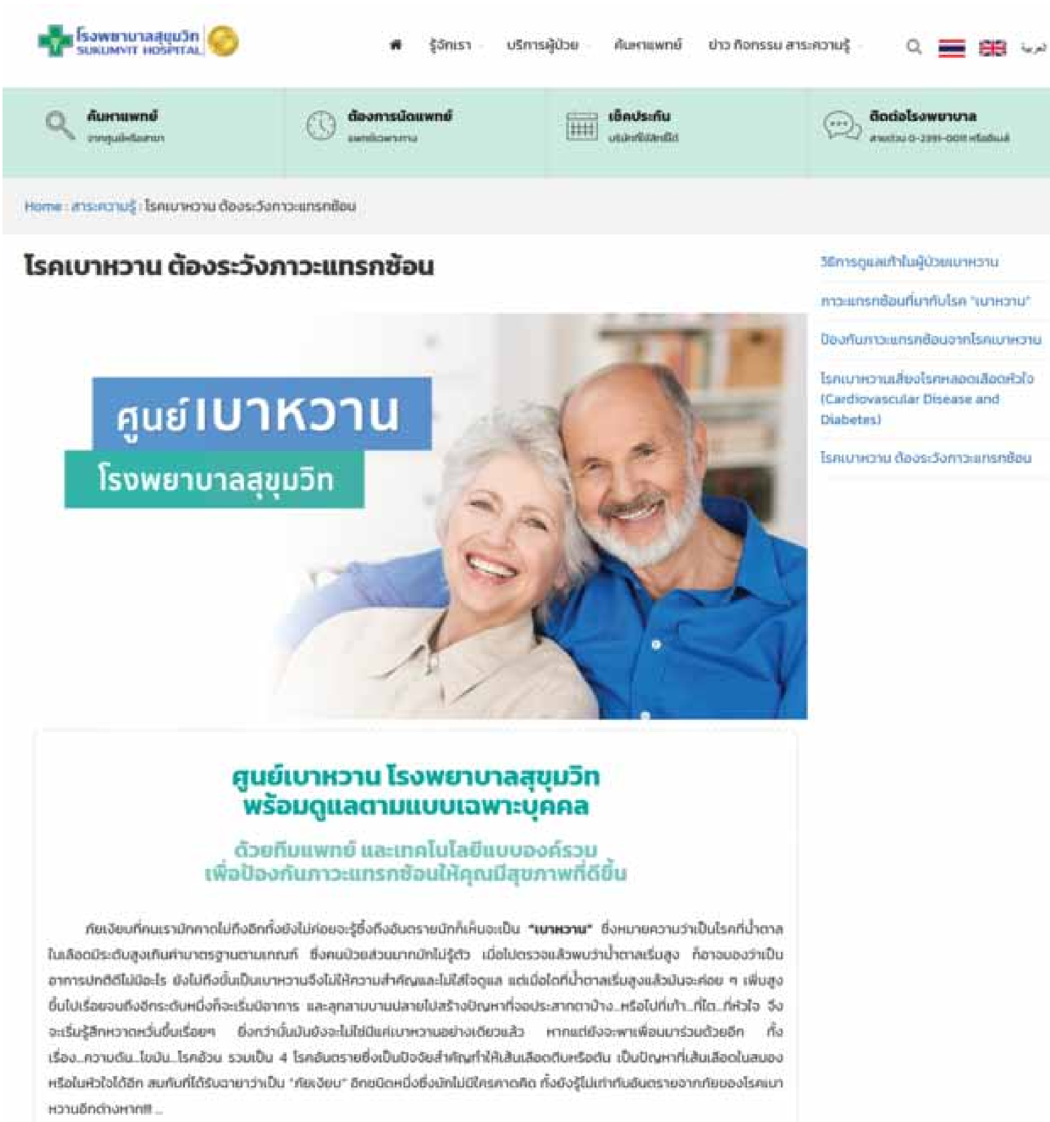
During the development of the researchers' baseline model in previous work (Saengkhunthod et al., 2021), they explored various methods to identify fake health and medical news as unreliable articles by identifying features using word cloud and ML to analyze unreliable medical articles. They collected samples of 297 reliable and 235 unreliable articles from seven websites, and analyzed the differences between them. The researchers tested the correlation between the variables, using χ^2 and information gain. The word cloud and term frequency-inverse frequency (TF-IDF) techniques were used to find reliable words. Then, they selected 20 features that affected article unreliability or reliability, and applied ML to classify the articles. The XGBoost was found to be the most effective, with an accuracy of 90.6% (Saengkhunthod et al., 2021). However, because the researchers' baseline model did not have a feature to indicate the purpose of the article, it could not distinguish between articles that had similar content but served different purposes, i.e., fake news (so-called unreliable articles), and not fake news (so-called reliable articles). Moreover, its training dataset had a limited number of counter-articles to defend or correct misunderstandings from unreliable articles on the same subject.

DATA COLLECTION

Based on the characteristics of health and medical websites described in the literature review by (Molina et al., 2021; Samuel & Zaiane, 2012; Zhou & Zafarani, 2020), the researchers selected the websites for data collection. Reliable health and medical articles were collected from accredited hospitals and well-known Thai healthcare websites ranked by the Internet Innovation Research Center (IIRC) (n.d). The articles were referenced to reliable sources and credible medical institutions. Unreliable health and medical articles were collected from websites that did not provide a reference to the source of the information, were not supervised by medical professionals, and the website developer was not named. The articles from websites shown in the Ministry of Digital Economy and Society (n.d.a.) were considered to be unreliable.

The researchers gathered 582 reliable and 435 unreliable articles (n = 1,017) (see Table 1). Richardson's Beautiful Soup (Richardson, 2021), a Python library package, was used to extract data from HTML and XML. The health and medical articles (so-called article) were divided into article data (Dataset A) and news data (Dataset B).

Figure 2. An Example of Reliable Data Collected From the Sukumvit Hospital Website (Rattanapraphat & Eawsinphanit, n.d)



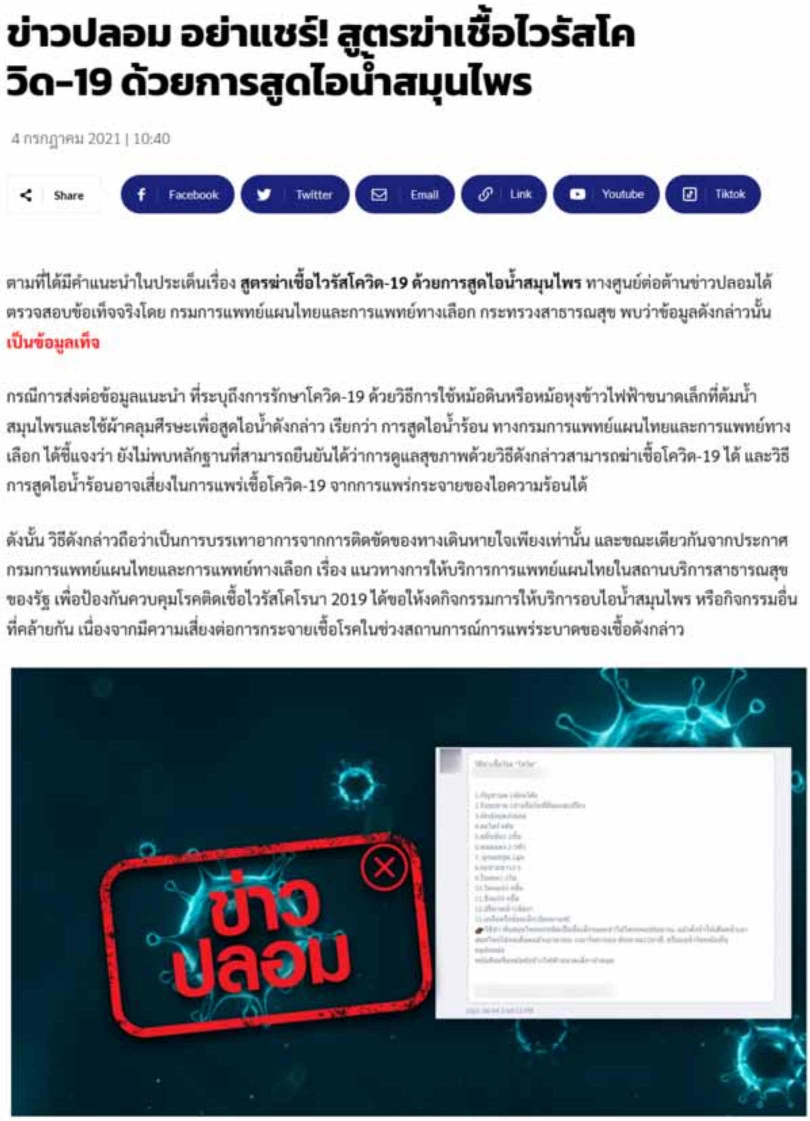
Dataset A

Dataset A was baseline data from public and private websites in the researchers' earlier model (Saengkhunthod et al., 2021). This dataset contained 617 articles, including 382 reliable and 235 unreliable articles. The researchers collected the title, content, and number of images from each website (see Figure 2). It included content about health from various viewpoints, such as herbs to cure diseases, education about various diseases, and solving health problems.

Dataset B

Dataset B contained news data from the Anti-Fake News Center website. The anti-fake news data contained news title, content explaining why it is fake, the name of the inspection agency, and an image of the fake news source, such as the Facebook site (see Figure 3). A web crawler was used to

Figure 3. An Example of Fake News Collected on the Anti-Fake News Site



ข่าวปลอม อย่าแชร์! สูตรฆ่าเชื้อไวรัสโควิด-19 ด้วยการสูดไอน้ำสมุนไพร

4 กรกฎาคม 2021 | 10:40

Share Facebook Twitter Email Link Youtube Tiktok

ตามที่ได้มีคำแนะนำในประเด็นเรื่อง **สูตรฆ่าเชื้อไวรัสโควิด-19 ด้วยการสูดไอน้ำสมุนไพร** ทางศูนย์ต่อต้านข่าวปลอมได้ตรวจสอบข้อเท็จจริงโดย กรมการแพทย์แผนไทยและการแพทย์ทางเลือก กระทรวงสาธารณสุข พบว่าข้อมูลดังกล่าวนั้น **เป็นข้อมูลเท็จ**

กรณีการส่งต่อข้อมูลแนะนำ ที่ระบุถึงการรักษาโควิด-19 ด้วยวิธีการใช้หม้อดินหรือหม้อหุงข้าวไฟฟ้าขนาดเล็กที่ต้มน้ำสมุนไพรและใช้ผ้าคลุมศีรษะเพื่อสูดไอน้ำดังกล่าว เรียกว่า การสูดไอน้ำร้อน ทางกรมการแพทย์แผนไทยและการแพทย์ทางเลือก ได้ชี้แจงว่า ยังไม่พบหลักฐานที่สามารถยืนยันได้ว่าการดูแลสุขภาพด้วยวิธีดังกล่าวสามารถฆ่าเชื้อโควิด-19 ได้ และวิธีการสูดไอน้ำร้อนอาจเสี่ยงในการแพร่เชื้อโควิด-19 จากการแพร่กระจายของไอความร้อนได้

ดังนั้น วิธีดังกล่าวถือว่าเป็นการบรรเทาอาการจากการติดขัดของทางเดินหายใจเพียงเท่านั้น และขณะเดียวกันจากประกาศกรมการแพทย์แผนไทยและการแพทย์ทางเลือก เรื่อง แนวทางการให้บริการการแพทย์แผนไทยในสถานบริการสาธารณสุขของรัฐ เพื่อป้องกันควบคุมโรคติดเชื้อไวรัสโคโรนา 2019 ได้ขอให้งดกิจกรรมการให้บริการอบไอน้ำสมุนไพร หรือกิจกรรมอื่นที่คล้ายกัน เนื่องจากมีความเสี่ยงต่อการกระจายเชื้อโรคในช่วงสถานการณ์การแพร่ระบาดของเชื้อดังกล่าว

crawl articles from the sources, and the researchers manually extracted fake news text in the images from the articles. Then, these extracted fake news were labeled data as unreliable article. On the other hand, the articles shown on the website were labeled as reliable articles. In total, the researchers collected 200 reliable and 200 unreliable (fake news) articles.

ANALYSIS OF THAI HEALTH AND MEDICAL ARTICLES ON WORD USAGE

Based on the literature review of Liu's study (2019), the researchers observed words frequently appearing in reliable and unreliable articles in Thai health and medical articles. They analyzed the data distribution and word cloud to identify the attributes of articles and the interquartile range of words about fruits, vegetables, herbs, disease, and organs.

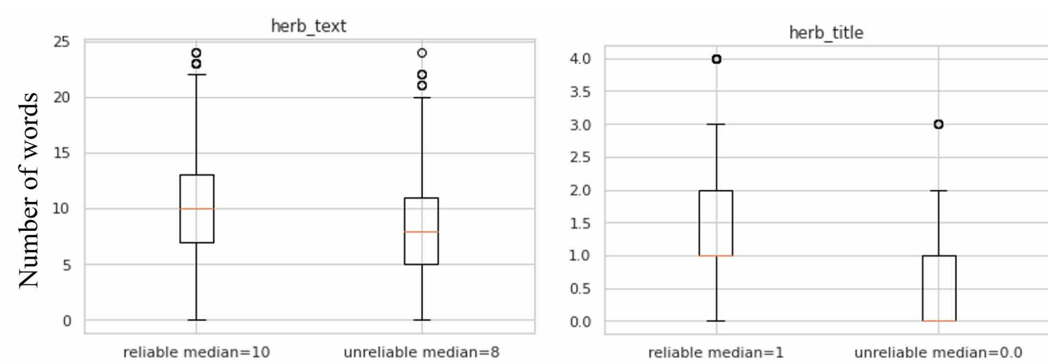
Table 1. Data Sources

Website	URL	Type of Website	Dataset	Number of Articles
Reliable Data Sources				
Medthai	https://medthai.com	Public website	A	118
Sukumvit Hospital	https://sukumvithospital.com	Hospital website	A	109
Med Mahidol	https://med.mahidol.ac.th	Hospital website	A	155
Anti-Fake News Center Thailand	https://www.antifakenewscenter.com/	Government site	B	200
Total Reliable				582
Unreliable Data Sources				
Dokkaew	https://dokkaew.wordpress.com	Public website	A	56
Bangpunsara	http://www.bangpunsara.com	Public website	A	53
Conloncancerzone	https://www.coloncancerzone.com	Public website	A	88
Eatonlinehealth	https://www.eatonlinehealth.com	Public website	A	38
Anti-Fake News Center Thailand	https://www.antifakenewscenter.com/	Government site	B	200
Total Unreliable				435

Herb-Related Words

Articles related to Thai health usually discuss herbs. Therefore, the researchers compared the interquartile range of these words in reliable and reliable articles (see Figure 4). The median number of herb-related words in reliable was 10 and in unreliable articles it was eight. The median number of herb-related words in titles of reliable articles was one and in unreliable articles, it was zero. The interquartile range of herb-related words in reliable articles was higher than that in unreliable articles, because reliable articles usually discussed multiple herbs and the articles were longer. Furthermore, herb-related words were significantly more frequent in the titles of reliable articles (see the right of Figure 4).

Figure 4. Number of Herb-Related Words in the Article Body (Left) and in the Article Title (Right)



The researchers found that the reliable articles contained a higher number of specific herb names than the unreliable articles (see Figure 5). In reliable articles, the largest (most frequent) words were กาแฟ (coffee), น้ำผึ้ง (honey), ชาเขียว (green tea), ตรีผลา (triphala), and ถั่งเช่า (cordyceps); whereas the most frequent words in unreliable articles were กาแฟ (coffee), ถั่วเหลือง (soybean), น้ำมันมะพร้าว (coconut oil), ย่านาง (bai-ya-nang), ฟ้าทะลายโจร (andrographis paniculata), and น้ำผึ้ง (honey) (see Figure 5). Therefore, the researchers used these words as new data features and took them through the feature selection process.

Fruit and Vegetable Words

The median number of fruit and vegetable words in reliable articles was seven, and in unreliable articles, it was five (see Figure 6). This may be because a reliable article aims to highlight the content of the article, so they need to be mentioned often, or they may be a longer article content (see Figure 6).

The researchers found that the reliable articles contained a higher number of specific fruit and vegetable names than the unreliable articles (see Figure 6). This may indicate that reliable articles want to emphasize the content of the article and explain in detail using repeated fruit and vegetable names. In reliable articles, the largest (most frequent) words were มะเขือเทศ (tomato), ส้ม (orange), กล้วยน้ำว้า (cultivated banana), มะกอก (olive), and พุทรา (jujube). On the other hand, unreliable articles tended to be shorter. The most frequently mentioned fruits and vegetables in unreliable articles were ส้ม (orange), มะเขือเทศ (tomato), กล้วยหอม (banana), เชอรี่ (cherry), แคนตาลูป (cantaloupe), ชมพู่ (rose apple), and บร็อคโคลี่ (broccoli) (see Figure 7). Therefore, the researchers used these words as new data features and took them through the feature selection process.

Figure 5. Word Clouds for Herbs in Reliable Articles (Left) and Unreliable Articles (Right)

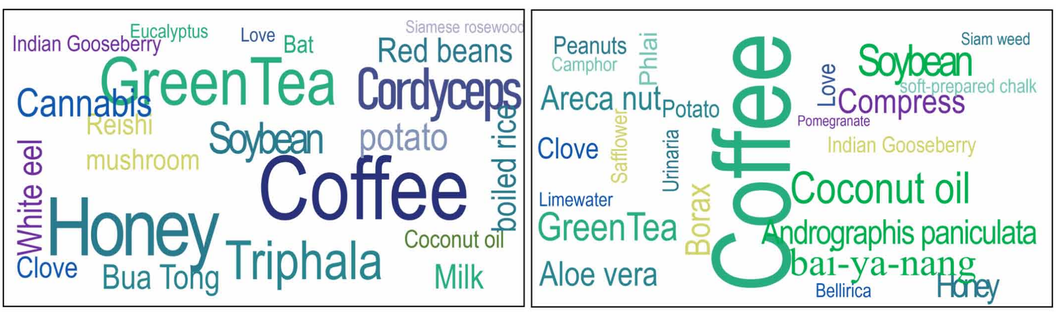


Figure 6. Number of Fruit and Vegetable Words in the Body of the Article

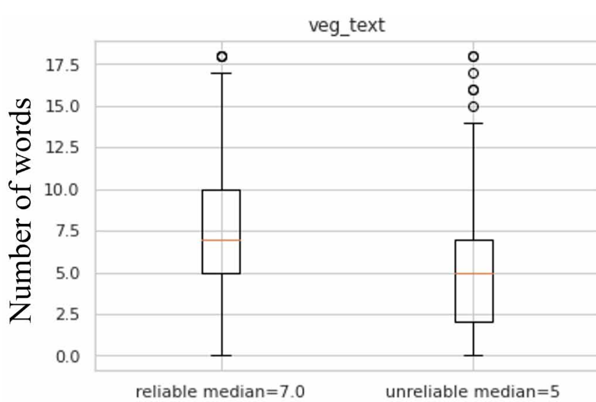


Figure 7. Word Clouds for Fruits, And Vegetables in Reliable Articles (Left) and Unreliable Articles (Right)

Disease and Organ-Related Words

Thai health and medical articles also frequently mention diseases and organs, such as cancer, diabetes, bone, and brain (see Figure 8). A list of diseases from the Medthai website (Medthai, n.d.) was used to filter only the names of diseases and organs. The median number of disease names and organ-related words in the body of reliable articles was nine, and in unreliable articles, it was five (see the left of Figure 10). Reliable articles mentioned more diseases and organs than unreliable ones in the body (see Figure 8).

The word cloud shows which diseases and organs were frequently mentioned in reliable and unreliable articles (see Figure 9). In unreliable articles, the largest (most frequent) words were มะเร็งปากมดลูก (cervical cancer), มะเร็งเต้านม (breast cancer), กระดูก (bone), มะเร็งตับ (liver cancer), ไต (kidney), สมอง (brain), ตา (eye), เบาหวาน (diabetes), ผม (hair), มะเร็งกระเพาะอาหาร (stomach cancer), มะเร็งต่อมลูกหมาก (prostate cancer), มะเร็งปอด (lung cancer), มะเร็งผิวหนัง (skin cancer), ฟัน (teeth), and กระ (freckles); whereas the most frequent words in reliable articles were เบาหวาน (diabetes), สมอง (brain), กระดูก (bone), หัวใจ (heart), ฟัน (teeth), ตา (eye), ผม (hair), ผิว (skin), and มะเร็งเต้านม (breast cancer), indicating a wider spread of different words in reliable articles (see Figure 9). This implies that unreliable articles tend to focus on a limited set of words, such as cancer, bone, diabetes, brain, and eyes, whereas reliable articles discuss a wider range of disease and organ-related words.

Figure 8. Number of Disease and Organ-Related Words in Article Bodies

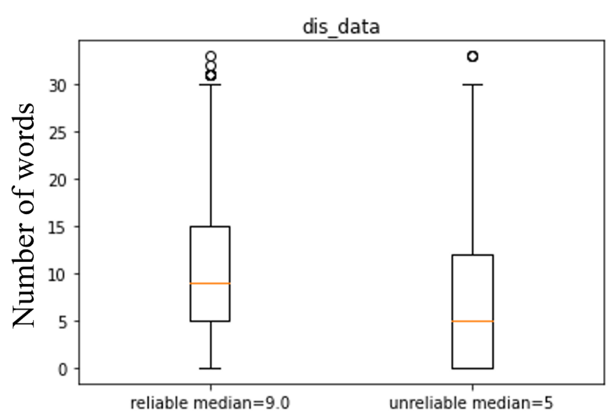
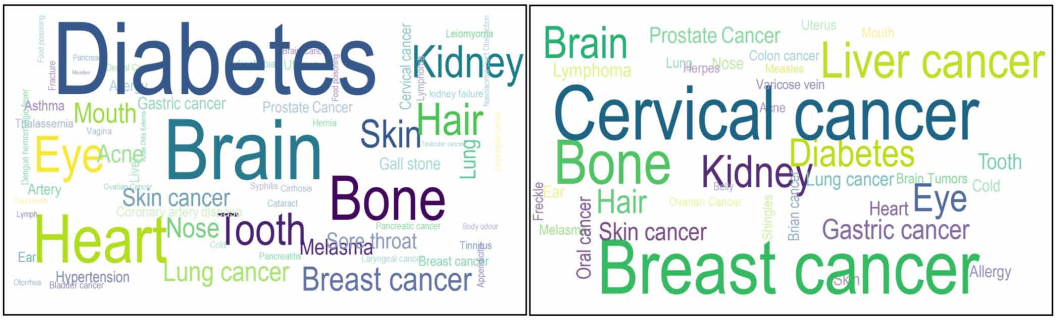


Figure 9. Word Clouds for Diseases and Organs That Appear in Reliable (Left) and Unreliable (Right) Articles

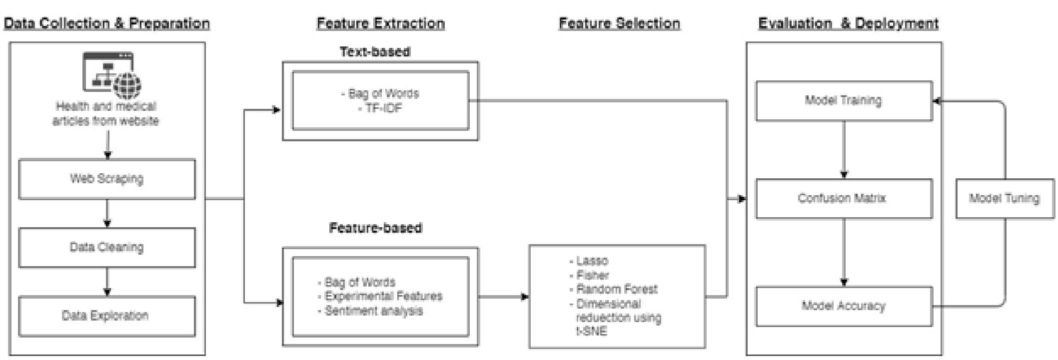


Based on the results from the word clouds (see Figures 7 and 9), the most frequent words in unreliable articles were selected as the experimental features because we wanted to focus on detecting unreliable data. This will detect unreliable data more accurately. Since certain words appeared in both reliable and unreliable articles, e.g., breast cancer, bone, brain, lime, and coffee (see Figures 7 and 9), and many reliable articles were created to defend or correct misunderstandings from unreliable articles on the same subject; these words cannot be used to reliably classify articles.

METHODOLOGY

The researchers collected health and medical articles using web scraping, cleaning, and then preprocessing for further steps. In the feature extraction process, they performed both text-based and feature-based methods (see Figure 10). In the text-based method, word cloud and TF-IDF techniques were applied to generate intra-text word features. The feature-based method was performed using bag-of-words techniques, statistical analysis for discovering new features, and sentiment analysis. Then, the features that most affected the performance of the model to detect unreliable health and medical articles were selected. After the feature selection process was completed, dimensionality reduction was performed using the t-distributed stochastic neighbor embedding (t-SNE) technique. Seven algorithms were used to train the model, including decision tree, k-nearest neighbors, logistic regression, naive Bayes, RF, support vector machine, and XGBoost. Finally, the confusion matrix was used to evaluate the performance of the ML models by assessing their precision, recall, accuracy, and F1 score.

Figure 10. Data Flow Process



Data Preprocessing

The articles on Thai websites were scraped using the Beautifulsoup library package, which produced different results due to the different structures of the websites. For example, some articles had a comment section, while others had a menu section that appeared in the article content, and some others contained special characters that were unnecessary. In the data cleaning process, the comment and menu sections that were unnecessary for classification were removed from the dataset of Medthai source. '\n' and '\t' were removed from the datasets of Bangpunsara and Dokkaew sources. As some words were written incorrectly in the researchers' dataset, PyThaiNLP package (Artificial Intelligence Research Institute of Thailand, 2019) wordnet module was used to find misspelled words and replace them with correct words. However, the output contained noise, such as place names, colloquial words, and misspelled words. Therefore, the researchers used WangchanBERTa, a pre-trained transformer based on the PyThaiNLP library (Artificial Intelligence Research Institute of Thailand, 2021). They used name entity recognition (NER) and tagging to define the part-of-speech of each word and cull out insignificant words, such as place and personal names. Then, the dataset was cleaned and transformed into a common format.

Feature Extraction

The feature extraction process mapped the pre-processed dataset into a new feature set. This process was divided into a text-based method for intra-text word features, and a feature-based method for numerical features.

Text-Based Method

The researchers used bag-of-words to convert a text into its numerical vector equivalent. The TF-IDF method was applied to find relevant words in the dataset. It enabled each word to be correlated with a number that indicated how important the word was in that document. The first step finds the term frequency, shown in Equation 1, is calculated from:

$$tf\left(term, document\right) = \frac{f\left(term, document\right)}{\sum_{term' \in document} f\left(term', document\right)} \tag{1}$$

The final step was TF-IDF, which is based on counting the number of documents (*N*) in the collection that contain the term (*t*) in question. The intuition is that a query term that occurs in many documents is not a good discriminator and should be given less weight than one that occurs in only a few documents. The heuristic to implement this is shown in Equation 2 (Stephen, 2004):

$$idf\left(term, allDocuments\right) = log\frac{N}{df\left(t\right)} \tag{2}$$

Feature-Based Method

The feature-based method transformed the value of each feature-based feature to lie in (0, 1), which is the max-min normalization shown in Equation 3, calculated by:

$$x^{'} = \frac{x - \min\left(x\right)}{\max\left(x\right) - \min\left(x\right)} \tag{3}$$

Figure 11. Word Cloud Created From Reliable Articles (Left) and Unreliable Articles (Right)

Features from the text-based method of data preprocessing are inputted into the feature selection process. On the other hand, the features from the feature-based method were extracted using word cloud, new features, and sentiment analysis.

1. World cloud

The researchers used the bag-of-words to create a word cloud that identified the most frequent words in the titles of reliable and unreliable articles, as illustrated in Figure 11. The articles also contain ‘อย่าแชร์ (do not share),’ ‘ข่าวปลอม (fake news),’ and, ‘บิดเบือน (distort),’ because we collected some anti-fake news articles that were created to prevent sharing of false information and discussion of misinformation (see the left side of Figure 11). Reliable and unreliable articles have some different words in their word clouds. The most common words found in unreliable articles were ‘มะเร็ง (cancer),’ ‘อาการ (symptom),’ and ‘กิน (eat)’ (see the right side of Figure 11).

2. Lexicon Features

The features generated from the result of section 4 are new features. They are characteristics of health and medical articles on Thai language websites, which contain herb, vegetable, and fruit: ÁÐ¹ÒÇ (lime), กระเทียม (garlic), น้ำมันมะพร้าว (coconut oil), มะระขี้นก (bitter gourd), มังคุด (mangosteen), กัญชา (cannabis), กาแฟ (coffee), น้ำผึ้ง (honey), มะละกอ (papaya), and มะขาม (tamarind); and disease and organ: มะเร็งปากมดลูก (cervical cancer), มะเร็งเต้านม (breast cancer), กระดูก (bone), มะเร็งตับ (liver cancer), ไต (kidney), สมอง (brain), ตา (eye), เบาหวาน (diabetes), ผม (hair), มะเร็งกระเพาะอาหาร (stomach cancer), มะเร็งต่อมลูกหมาก (prostate cancer), มะเร็งปอด (lung cancer), มะเร็งผิวหนัง (skin cancer), ฟัน (teeth), and กระ (freckles).

3. Sentiment-Based Features

The researchers first performed sentiment analysis in word cloud with TF-IDF to find words that could distinguish reliable and unreliable articles. The result of TF-IDF revealed that “ไม่จริง (untrue),” “อย่าแชร์ (do not share),” and และ “ไม่เชื่อ (do not believe)” were negative words in reliable articles. On the other hand, ดีจริง ๆ (really good),” and ของดี (good stuff),” which are positive words, were found in unreliable articles. This is because unreliable articles are often written to support and agree with those medical beliefs to convince them to believe in one another. On the other hand, reliable articles often contain content about how to disprove fake news on the Internet or attack those false articles in order to prevent people from believing those false articles. This leads the researchers to the hypothesis that reliable articles use negative words, and unreliable articles use positive words.

Table 2. Proposed Feature Set From Feature Extractions of Feature-Based Method

Feature Extractions	Feature Set	Number of Features
Word Cloud	"มะเร็ง (cancer)," "กิน (eat)," "รักษา (cure)," "อาการ (symptom)," "สมุนไพร (herb)," "โควิด (Covid)," "โซดา (soda)"	7
Experimental Features	List of herb, vegetable, and fruit: มะนาว (lime), กระเทียม (garlic), น้ำมันมะพร้าว (coconut oil), มะระขี้นก (bitter gourd), มังคุด (mangosteen), กัญชา (cannabis), กาแฟ (coffee), น้ำผึ้ง (honey), มะละกอ (papaya) and มะขาม (tamarind). List of disease and organ: มะเร็งปากมดลูก (cervical cancer), มะเร็งเต้านม (breast cancer), กระดูก (bone), มะเร็งตับ (liver cancer), ไต (kidney), สมอง (brain), ตา (eye), เบาหวาน (diabetes), ผม (hair), มะเร็งกระเพาะอาหาร (stomach cancer), มะเร็งต่อมลูกหมาก (prostate cancer), มะเร็งปอด (lung cancer), มะเร็งผิวหนัง (skin cancer), ฟัน (teeth), and กระ (freckles)	25
Sentiment Analysis		1

The researchers therefore assigned a label to Dataset B, assigning negative labels as one (1) to reliable articles, and positive label as zero (0) to unreliable articles for classification to prove the hypothesis they set. The TF-IDF technique was used to transform text into matrix format to be used as a feature in classification. The researchers split the data into 70% as training and 30% as a testing set. Then, they used logistic regression to classify the labeled Dataset B. The classification accuracy was 99%, indicating that sentiment analysis can be used to classify reliable and unreliable articles. They then experimented with adding features from the sentiment analysis result to predict unreliable data. This was performed to test whether adding such a feature could increase the accuracy of classification. Therefore, this sentiment analysis model was used to analyze the sentiment in Dataset A.

The researchers obtained 33 numerical features; seven features were selected from the word cloud method, 25 features from the proposed new features, and one from the sentiment analysis (see Table 2).

Feature Selection

The researchers used the LASSO parameter estimation method, the Fisher score, and the RF method to select the best features for the model classifying the fake news in datasets, and then compared the results from each method.

LASSO

Tibchirani (1996) proposed a parameter estimation method called LASSO that represents the least absolute selection and shrinkage operator. The objective was to reduce the shape and select the features simultaneously. LASSO minimizes the residual sum of squares, subject to the sum of the absolute value of the coefficient being less than a constant. If the coefficient is greater than zero, then the independent variable is correlated with the dependent variable.

Fisher

The key idea of the Fisher score is to find a subset of features such that in the data space spanned by the selected features, the distances between data points in different classes are as large as possible, while the distances between data points in the same class are as small as possible (Gu, Li, & Han, 2012).

RF

The RF method is adept at identifying relevant features with only slight main effects in high-dimensional data (Reif et al., 2006).

The researchers reduced the dimension of the data to visualize, examine, and find the best features using t-SNE technique (van der Maaten & Hinton, 2008). Since their datasets were nonlinear, the t-SNE algorithm was applied to perform nonlinear dimensional reduction, because it can discover a way to project the data into a lower-dimensional space/embedding while preserving the original high-dimensional clustering (van der Maaten and Hinton, 2008). Moreover, it is usually more important to keep low-dimensional representations of very similar data points closer together, which is typically not possible with linear mapping (van der Maaten & Hinton, 2008).

Modeling

We used five machine-learning algorithms (Decision Tree, RF, XGBoost, naïve Bayes, support vector machine, logistic regression, and k-nearest neighbors) from our previous baseline model (Saengkhunthod et al., 2021) and two additional algorithms, the RF and naive Bayes, were added. Gridsearchcv, a function of Scikit-learn (Pedregosa et al., 2011), was used to fine-tune their ML models to find the parameter to obtain the best accuracy for each model and prevent overfitting (Liashchynskyi & Liashchynskyi, 2019).

RF

RF is an ensemble learning method that creates multiple decision trees during training (Breiman, 2001). It uses a bagging technique to randomize the unique dataset and set of features to each tree, and then conducts a majority vote to select the class of prediction for each tree.

Naive Bayes

Naive Bayes is a classification algorithm that uses Bayes' theorem with the assumption of independence of variables (Murty & Devi, 2011). Naive Bayes is a scalable classification method that requires linear learning by applying the principle of probability and decision-making rules, known as the MAP decision rule.

XGBoost

XGBoost is a learning system for tree boosting that was developed from a gradient boost. XGBoost uses a decision tree as a weak predictor. Each decision tree learns from the previous error tree, thus improving the prediction accuracy. The learning of the tree continues until it is sufficiently deep, and the model stops learning when no pattern of errors from the previous tree is left to learn (Chen & Guestrin, 2016).

kNN

The kNN algorithm is a lazy learning classification algorithm. It finds the nearest distance between the points and conducts a majority vote to predict the class of each point (Guo et al., 2003).

SVM

The main function of the SVM algorithm is to find a hyperplane in an N-dimensional space. N is the number of features that distinctly classifies the data points (Evgeniou & Pontil, 1999). The best hyperplane that can separate classes of data points was selected. Maximizing margins provides some reinforcement to future data points and can be classified more confidently.

Logistic Regression

Logistic regression is a statistical machine-learning model that uses a logistic function to classify the binary dependent variable (Tolles & Meurer, 2016). This algorithm outputs the probability of predicting the class of data when the input of the model is a numerical feature.

Decision Tree

A decision tree is a classification model that uses a tree-based model (Song & Ying, 2015). This model selects the most related feature as the root node, and then selects the next node by computing the information gain. The higher the information gain values of the feature, the more the features are related to the data.

EXPERIMENTAL RESULTS AND DISCUSSION

Result of Dimensionality Reduction

The researchers applied the t-SNE technique to reduce the dimensions of the data and created a scatter plot to discuss the distribution of data selected by each feature selection method. The results from three feature selections are illustrated in Figures 12 through 14. The model was able to separate reliable and unreliable articles. LASSO and Fisher (see Figures 12 and 13) use straight lines to divide the data into two groups, and this ensured that the modeling process was continued.

Results of Feature Selection and Model Evaluation

Each dataset was split into a training set of 70% and a testing set of 30% of the total data. The detection model using XGBoost outperformed among others in all performance matrices. Specifically, XGBoost with LASSO and Random feature selections were the most effective, with 97.66% accuracy, an improvement of 7% from the previous model (Saengkhunthod et al., 2021).

The researchers summarized all the feature selections to select the set of features that provided the most accurate prediction, by comparing the average precision, recall, and accuracy (see Table 4). LASSO provided the highest precision and accuracy, and RF gave the highest recall and F1 score.

Figure 12. Grouping of the Data After Dimensionality Reduction by LASSO Feature Selection

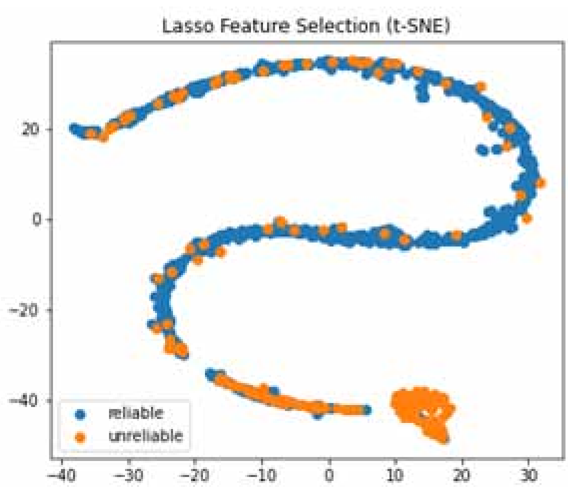


Figure 13. Grouping of the Data After Reduction in Dimensionality by Fisher Feature Selection

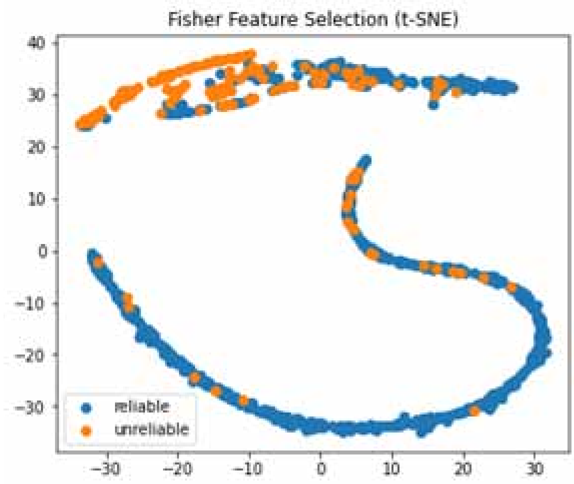
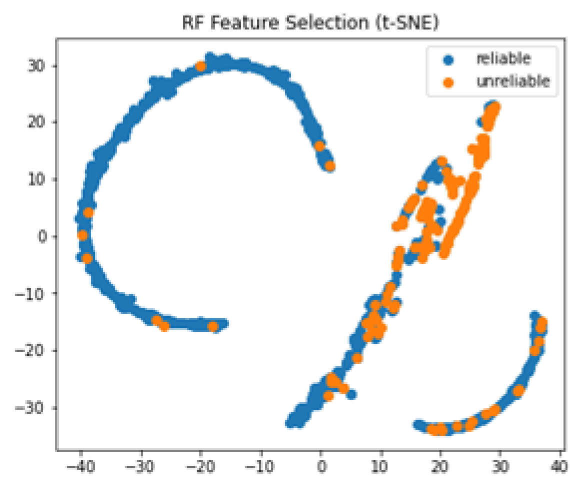


Figure 14. Grouping of the Data After Dimensionality Reduction by RF Feature Selection



The researchers used the receiver operating characteristic (ROC) curve to measure the performance of the models to help with the decision as to the best model. They created three ROC curves: a feature selection model using the LASSO technique, the RF technique, and the Fisher technique, respectively.

As can be seen in Figure 15, the ROC curve of feature-selection model using the LASSO technique presents that the XGBoost model approaches one (1) more closely than those of the RF and Logistic Regression, respectively.

The ROC curve of the feature-selection model using the RF technique presents that the XGBoost model approaches one (1) more closely than those of RF and Logistic Regression, respectively (shown in Figure. 16).

Figure 17 illustrates the ROC curve of the feature-selection model using the Fisher technique, showing that the XGBoost model approaches one (1) more closely than those of the RF and Decision Tree, respectively.

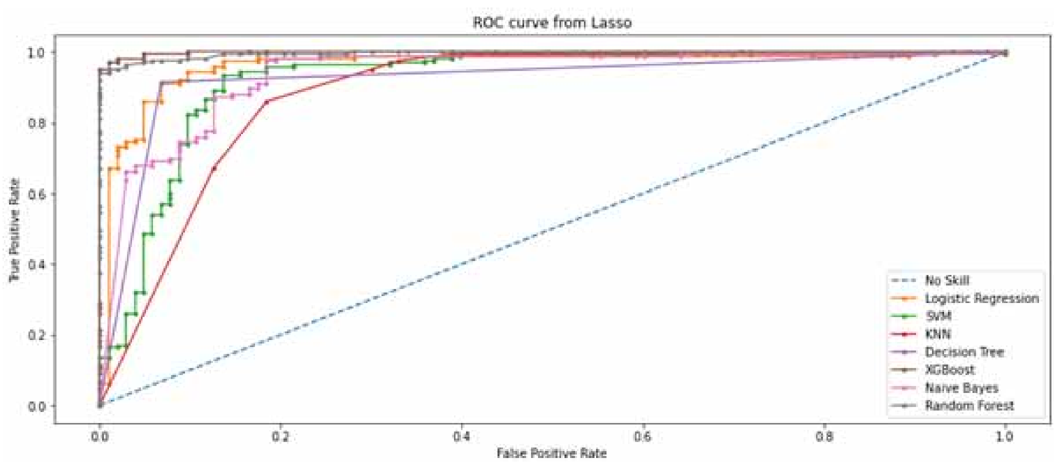
Table 3. Performance Matrix of Feature Selection Methods in Seven Models

Algorithm/Feature Selection	Precision			Recall			F1 Score			Accuracy		
	LASSO	RF	Fisher	LASSO	RF	Fisher	LASSO	RF	Fisher	LASSO	RF	Fisher
Decision Tree	89.19	88.87	91.04	89.95	90.48	90.6	89.53	89.64	90.81	92.64	92.64	93.65
kNN	81.83	86.98	86.98	76.3	79.62	78.62	78.49	82.47	82.47	86.29	88.96	88.96
Logistic Regression	87.18	88.54	83.48	81.64	84.63	84.18	83.94	86.35	83.82	89.63	90.97	88.63
Naïve Bayes	82.8	78.77	82.23	75.73	73.25	74.76	77.1	74.32	76.08	80.22	74.32	79.48
RF	96.62	95.9	95.64	95.62	95.41	94.66	96.11	95.65	95.14	97.32	96.99	96.66
SVM	90.42	88.19	86.47	89.11	88.99	88.34	89.74	88.58	87.35	92.98	91.97	90.97
XGBoost	**96.87**	**96.87**	**97.32**	**96.37**	**96.37**	**96.62**	**96.62**	**96.62**	**95.62**	**97.66**	**97.66**	**97.32**

Table 4. Average Result of Each Feature Selection

Feature Selection	Precision	Recall	F1 Score	Accuracy
LASSO	**89.27**	86.39	87.36	**90.96**
RF	89.16	**86.96**	**87.66**	90.50
Fisher	89.02	86.83	87.33	90.81

Figure 15. ROC Curve Result of Feature-Selection Models Using LASSO Technique



From Figures 15-17, it can be concluded that the ROC curves of the feature-selection models using LASSO closest to one (1) are the XGBoost model, the RF model, and the Logistic Regression model, respectively.

The researchers conducted the experiments by matching their datasets with two sets of features i.e., baseline (20 features from (Saengkhunthod et al., 2021) and proposed (33 features, listed in Table 2) to create three different models (see Table 5). Model#1, with baseline features trained with Dataset A and tested with Dataset B, had an accuracy of 85.82%. This implied that Model#1 was unable to predict fake news dataset, which contains similar content, but serves different purposes. The researchers used 70% of the Datasets (A+B) for training and 30% for testing Model#2 with baseline

Figure 16. ROC Curve Result of Feature-Selection Models Using RF Technique

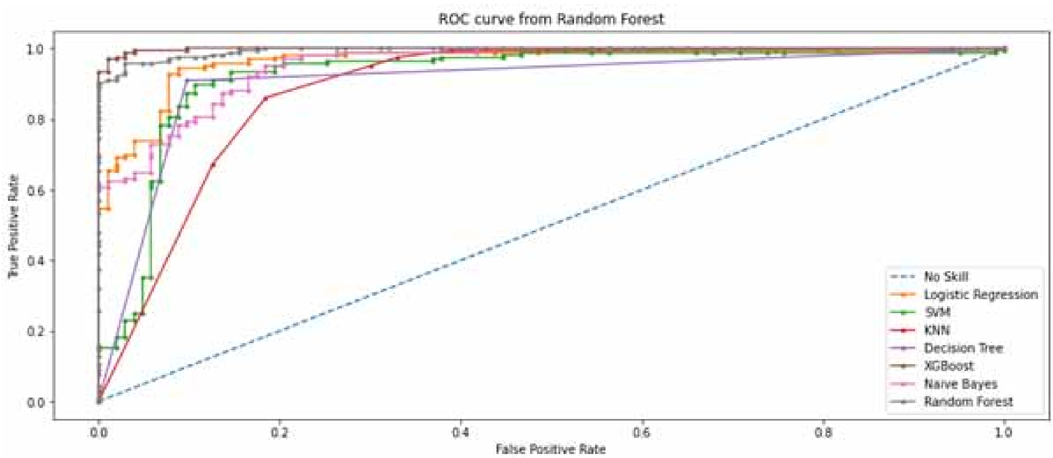
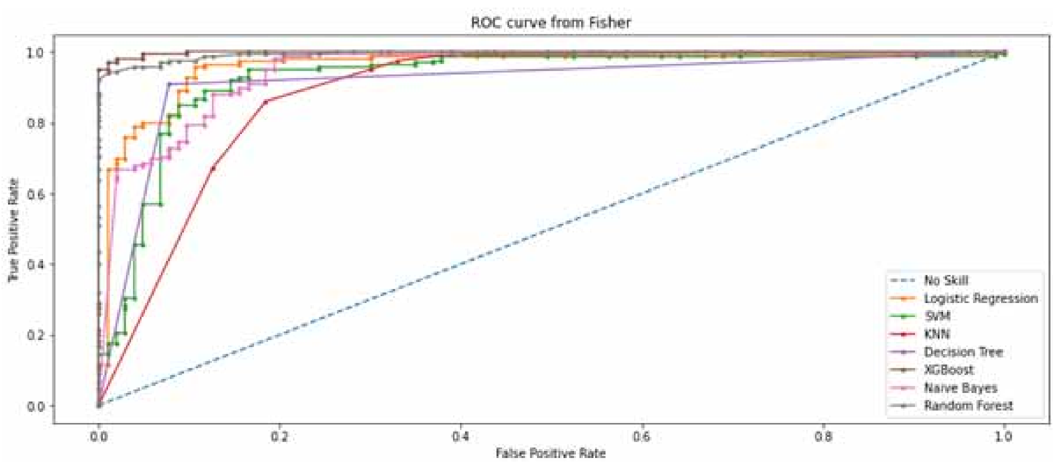


Figure 17. ROC Curve Result of Feature-Selection Models Using the Fisher Technique



features and Model#3. The accuracy of Model#2 was 94.78%. Remarkably, Model#3 with additional features proposed from the LASSO feature selection process achieved the best performance with an accuracy of 97.76%. Compared to the performance of Model#2, this was an increase of almost 3% by using LASSO feature selection and the proposed features in the XGBoost model.

Therefore, Model#3 revealed that the addition of Dataset B and the proposed features based on feature selection significantly increased accuracy by up to 7.16% (from 90.60% to 97.76%) and helped manage the problem of detecting articles that contained similar content, but that serve different purposes. This suggests that sentiment analysis can be a useful tool to help detect fake news, a finding consistent with a study by Ajao at el. (2019). However, the accuracy of the researchers' improved model was not as high as that of the model by Aphiwongsophon and Chongstitvatana (2018), which used different datasets. Aphiwongsophon and Chongstitvatana (2018) used short-text and Tweet data, and feature selection was not used to analyze the most relevant features. Compared with Liu et al.'s (2019) research, the researchers fine-tuned the model using the Gridsearchcv library (van der Maaten & Hinton, 2008) to find the parameters that provide the model with the highest predictive accuracy.

Table 5. Detail of Experiments

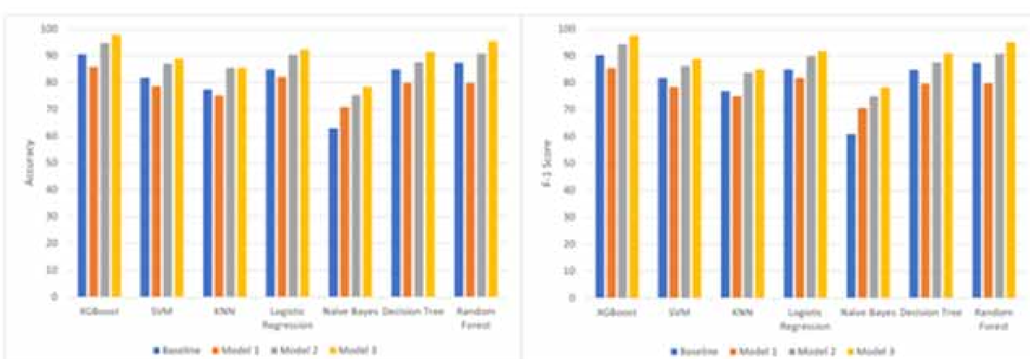
Model	Experimental Purpose	Training Dataset	Test Dataset	Feature	Feature Selection	Accuracy	F1 Score
Baseline [15]	To make the baseline model learn and assess its performance using only health and medical articles and 20 features.	A	A	Baseline	None	90.60	90.4
1	To assess the performance of baseline model with the news dataset, which contains similar content, but serves different purposes.	A	B	Baseline	None	85.82	86.46
2	To make the baseline model learn and assess its performance using two datasets.	A+B (70%)	A+B (30%)	Baseline	None	94.78	94.39
3	To improve performance of the baseline model and solve the problem of detecting fake news in similar content, but different objectives.	A+B (70%)	A+B (30%)	Baseline + Proposed	LASSO	**97.76**	**97.54**

A comparison of the performances of all models based on LASSO feature selection and proposed features is shown in Figure 18. The Model#3 showed an outstanding performance on accuracy and the F1 score of the XGBoost technique.

CONCLUSION AND DISCUSSION

The primary objectives of this study were to improve the performance of the researchers' proposed machine-learning detection model of fake news in health and medical articles on Thai websites, and to solve the problem of detecting fake news in content that is similar, but has different objectives. Their new and improved model, with 33 features (one text-based, 25 new features, and one sentimental analysis feature), can accurately classify unreliable (fake news) and reliable (not fake news) health-related articles. Specifically, sentiment analysis can help identify similar articles with different

Figure 18. Performances of All Models: Accuracy (Left) And F1 Score (Right)



objectives from two datasets, including the dataset from Saengkhunthod et al.'s (2021) research and a new dataset from a fake news source. The most accurate predictive model was XGBoost with LASSO selection, with a 97.76% accuracy, a 7.16% improvement from their earlier model (Saengkhunthod et al., 2021). Compared with the deep learning models of Nadeem et al. (2023) and Maity et al. (2023), the researchers' ML model requires less computational cost, which is suitable for Thai organizations that have limitations in computing hardware and/or lack budgets for deep learning infrastructures. This model would also be beneficial for anyone who would like to adopt the model for further implementation. For example, the Anit-Fake News Center of Thailand can implement the model for an automatic fake news detection system, which would reduce the time and speed up detection task.

There are many opportunities for future work. Since this research focused only on words, punctuation marks and other patterns of fake news articles, such as irrelevant medical images, should be considered to construct a model to detect other patterns of unreliable health and medical articles. In addition, the most accurate predictive model, XGBoost with LASSO, can be further analyzed on the Thai fake news dataset, such as LimeSoda dataset (Payoungkhamdee et al., 2021), and other health websites written in Tai-Kadai languages, such as Lao language, because Thai language is classified as a Tai-Kadai language (Diller, Edmondson, & Luo, 2004).

CONFLICTS OF INTEREST

We wish to confirm that there are no known conflicts of interest associated with this publication and there has been no significant financial support for this work that could have influenced its outcome.

FUNDING STATEMENT

This work was funded by the King Mongkut's Institute of Technology Ladkrabang Research Fund, Grant number 2564-02-06-001.

PROCESS DATES

Received: 1/8/2024, Revision: 3/4/2024, Accepted: 3/22/2024

CORRESPONDING AUTHOR

Correspondence should be addressed to Kanokwan Atchariyachanvanich (Thailand, kanokwan@it.kmitl.ac.th)

REFERENCES

Ajao, O., Bhowmik, D., & Zargari, S. (2019, May). Sentiment aware fake news detection on online social networks. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 2507-2511). IEEE. oi:10.1109/ICASSP.2019.8683170

Aphiwongsophon, S., & Chongstitvatana, P. (2018, July). Detecting fake news with machine learning method. In *2018 15th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)* (pp. 528-531). IEEE. oi:10.1109/ECTICon.2018.8620051

Artificial Intelligence Research Institute of Thailand. (2019). *PyThaiNLP resources.* (in Thai). https://github.com/PyThaiNLP/pythainlp

Artificial Intelligence Research Institute of Thailand. (2021). pythainlp.*wangchanberta.* (in Thai). https://pythainlp.github.io/dev-docs/api/wangchanberta.html

Aslam, N., Khan, I., Alotaibi, F. S., Aldaej, L. A., & Aldubaikil, A. K. (2021). Fake detect: A deep learning ensemble model for fake news detection. *Complexity*, *2021*, 1–8. Advance online publication. doi:10.1155/2021/5557784

Barve, Y., & Saini, J. R. (2023). Detecting and classifying online health misinformation with 'Content Similarity Measure (CSM)' algorithm: An automated fact-checking-based approach. *The Journal of Supercomputing*, *79*(8), 9127–9156. doi:10.1007/s11227-022-05032-y PMID:36644509

Barve, Y., Saini, J. R., Pal, K., & Kotecha, K. (2022). A novel evolving sentimental bag-of-words approach for feature extraction to detect misinformation. [IJACSA]. *International Journal of Advanced Computer Science and Applications*, *13*(4), 266–275. doi:10.14569/IJACSA.2022.0130431

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. doi:10.1023/A:1010933404324

Chen, T., & Guestrin, C. (2016, August). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785-794). oi:10.1145/2939672.2939785

Dey, A., Rafi, R. Z., Parash, S. H., Arko, S. K., & Chakrabarty, A. (2018, June). Fake news pattern recognition using linguistic analysis. In *2018 Joint 7th International Conference on Informatics, Electronics & Vision (ICIEV) and 2018 2nd International Conference on Imaging, Vision & Pattern Recognition (icIVPR)* (pp. 305-309). IEEE. oi:10.1109/ICIEV.2018.8641018

Diller, A., Edmondson, J., & Luo, Y. (2004). *The Tai-Kadai Languages* (1st ed.). Routledge., doi:10.4324/9780203641873

ETDA (Electronic Transactions Development Agency). (2020). *Survey of Internet usage behavior in Thailand 2020*. https://www.etda.or.th/th/newsevents/pr-news/ETDA-released-IUB-2020.aspx

Evgeniou, T., & Pontil, M. (1999). Support vector machines: Theory and applications. In *Advanced course on artificial intelligence* (pp. 249–257). Springer Berlin Heidelberg., doi:10.1007/3-540-44673-7_12

Fake-news. (n.d.a). Fake news. *Cambridge Dictionary online.* https://dictionary.cambridge.org/dictionary/english/fake-news

Fake-news. (n.d.b). Fake news. *Collins Dictionary online.* https://www.collinsdictionary.com/dictionary/english/fake-news

Guo, G., Wang, H., Bell, D., Bi, Y., & Greer, K. (2003). KNN model-based approach in classification. In *On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE: OTM Confederated International Conferences, CoopIS, DOA, and ODBASE 2003, Catania, Sicily, Italy, November 3-7, 2003. Proceedings* (pp. 986-996). Springer Berlin Heidelberg. doi:10.1007/978-3-540-39964-3_62

IIRC (Internet Innovation Research Center). (n.d.). *Thailand Web Directory and Advance Web Statistics at Truehits.net*. https://truehits.net/script/r.php?id=7

Iwendi, C., Mohan, S., Ibeke, E., Ahmadian, A., & Ciano, T. (2022). Covid-19 fake news sentiment analysis. *Computers & Electrical Engineering*, *101*, 107967. doi:10.1016/j.compeleceng.2022.107967 PMID:35474674

Jadhav, P., & Shukla, R. K. (2024). Deep learning analysis for revealing fake news using linguistic complexity and semantic signatures. *International Journal of Intelligent Systems and Applications in Engineering*, *12*(12s), 458–465.

Liashchynskyi, P., & Liashchynskyi, P. (2019). Grid search, random search, genetic algorithm: A big comparison for NAS. ArXiv, abs/1912.06059

Liu, Y., Yu, K., Wu, X., Qing, L., & Peng, Y. (2019). Analysis and detection of health-related misinformation on Chinese social media. *IEEE Access : Practical Innovations, Open Solutions*, *7*, 154480–154489. doi:10.1109/ACCESS.2019.2946624

Maity, K., Bhattacharya, S., Phosit, S., Kongsamlit, S., Saha, S., & Pasupa, K. (2023). HANCaps: A two-channel deep learning framework for fake news detection in Thai. In *International Conference on Neural Information Processing* (pp. 204-215). Singapore: Springer Nature Singapore.

MDES (Ministry of Digital Economy and Society). (n.d.a.) *Anti-Fake News Center Thailand*. https://www.antifakenewscenter.com

MDES (Ministry of Digital Economy and Society). (n.d.b) Fake news Don't share! lime soda cures cancer (¢èÒÇ»ÅÍÁ ÍÂèÒá^ªÃì! ÁÐ¹ÒÇâ«´ÒÃÑ¡ÉÒâÃ¤ÁÐàÃç§)." https://bit.ly/3lvQGVD

Medthai (n.d.). Popular and featured posts. Medthai. https://medthai.com

Molina, M. D., Sundar, S. S., Le, T., & Lee, D. (2021). "Fake news" is not simply false information: A concept explication and taxonomy of online content. *The American Behavioral Scientist*, *65*(2), 180–212. doi:10.1177/0002764219878224

Mookdarsanit, P., & Mookdarsanit, L. (2021). The covid-19 fake news detection in Thai social texts. *Bulletin of Electrical Engineering and Informatics*, *10*(2), 988–998. doi:10.11591/eei.v10i2.2745

Murty, M. N., & Devi, V. S. (2011). *Pattern Recognition: An Algorithmic Approach*. Springer Science & Business Media. doi:10.1007/978-0-85729-495-1

Nadeem, M. I., Mohsan, S. A. H., Ahmed, K., Li, D., Zheng, Z., Shafiq, M., Karim, F. K., & Mostafa, S. M. (2023). HyproBert: A fake news detection model based on deep hypercontext. *Symmetry*, *15*(2), 296. doi:10.3390/sym15020296

Nyow, N. X., & Chua, H. N. (2019, November). Detecting fake news with tweets' properties. In *2019 IEEE Conference on Application, Information and Network Security (AINS)* (pp. 24-29). IEEE. oi:10.1109/AINS47559.2019.8968706

Ozbay, F. A., & Alatas, B. (2020). Fake news detection within online social media using supervised artificial intelligence algorithms. *Physica A*, *540*, 123174. doi:10.1016/j.physa.2019.123174

Palani, B., & Elango, S. (2023). BBC-FND: An ensemble of deep learning framework for textual fake news detection. *Computers & Electrical Engineering*, *110*, 108866. doi:10.1016/j.compeleceng.2023.108866

Payoungkhamdee, P., Porkaew, P., Sinthunyathum, A., Songphum, P., Kawidam, W., Loha-Udom, W., Boonkwan, P., & Sutantayawalee, V. (2021, December). LimeSoda: Dataset for fake news detection in healthcare domain. In *2021 16th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP)* (pp. 1–6). IEEE. doi:10.1109/iSAI-NLP54397.2021.9678187

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*(85), 2825–2830. http://arxiv.org/abs/1201.0490

Rattanapraphat, T., & Eawsinphanit, V. (n.d.). *Diabetes need to be careful of complications*. (in Thai). https://sukumvithospital.com/healthcontent.php?id=3499

Reif, D. M., Motsinger, A. A., McKinney, B. A., Crowe, J. E., & Moore, J. H. (2006, September). Feature selection using a random forests classifier for the integrated analysis of multiple data types. In *2006 IEEE Symposium on Computational Intelligence and Bioinformatics and Computational Biology* (pp. 171-178). IEEE. oi:10.1109/CIBCB.2006.330987

Reis, J. C., Correia, A., Murai, F., Veloso, A., & Benevenuto, F. (2019). Supervised learning for fake news detection. *IEEE Intelligent Systems*, *34*(2), 76–81. doi:10.1109/MIS.2019.2899143

Richardson, L. (2021). *Beautiful Soup Documentation.* Beautiful Soup 4.12.0. documentation. https://www.crummy.com/software/BeautifulSoup/bs4/doc/

Saengkhunthod, C., Kerdnoonwong, P., & Atchariyachanvanich, K. (2021, January). Detection of unreliable medical articles on Thai websites. In *2021 13th International Conference on Knowledge and Smart Technology (KST)* (pp. 102-107). IEEE. oi:10.1109/KST51265.2021.9415756

Samuel, H. W., & Zaiane, O. R. (2012, January). PSST... privacy, safety, security, and trust in health information websites. In *Proceedings of 2012 IEEE-EMBS International Conference on Biomedical and Health Informatics* (pp. 584-587). IEEE. oi:10.1109/BHI.2012.6211650

Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media: A data mining perspective. *SIGKDD Explorations*, *19*(1), 22–36. doi:10.1145/3137597.3137600

Singh, M. K., Ahmed, J., Alam, M. A., Raghuvanshi, K. K., & Kumar, S. (2023). A comprehensive review on automatic detection of fake news on social media. *Multimedia Tools and Applications*, *83*(16), 1–34. doi:10.1007/s11042-023-17377-4

Song, Y. Y., & Ying, L. U. (2015). Decision tree methods: Applications for classification and prediction. *Shanghai Jingshen Yixue*, *27*(2), 130–135. doi:10.11919/J.ISSN.1002-0829.215044 PMID:26120265

Songram, P. (2019). Detection of unreliable and reliable pages on Facebook. *Artificial Life and Robotics*, *24*(2), 278–284. doi:10.1007/s10015-018-0509-z

Stephen, E. (2004). Understanding inverse document frequency: On theoretical arguments for IDF. *The Journal of Documentation*, *60*(5), 503–520. doi:10.1108/00220410410560582

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, *58*(1), 267–288. doi:10.1111/j.2517-6161.1996.tb02080.x

Tolles, J., & Meurer, W. J. (2016). Logistic regression: Relating patient characteristics to outcomes. *Journal of the American Medical Association*, *316*(5), 533–534. doi:10.1001/jama.2016.7653 PMID:27483067

Treharne, T., & Papanikitas, A. (2020). Defining and detecting fake news in health and medicine reporting. *Journal of the Royal Society of Medicine*, *113*(8), 302–305. doi:10.1177/0141076820907062 PMID:32780974

Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, *9*(11), 2579–2605.

Wu, L., Morstatter, F., Carley, K. M., & Liu, H. (2019). Misinformation in social media: Definition, manipulation, and detection. *SIGKDD Explorations*, *21*(2), 80–90. doi:10.1145/3373464.3373475

Zhou, X., & Zafarani, R. (2020). A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys*, *53*(5), 1–40. doi:10.1145/3395046

Kanokwan Atchariyachanvanich is an Assistant Professor at School of Information Technology, King Mongkut's Institute of Technology Ladkrabang, Thailand. She received a B.Sc. in Information Technology from Assumption University, a M.S. in Information Management from the Asian Institute of Technology, a M.M in Public Administration from Tsinghua University and a Ph.D. from the Graduate University for Advanced Studies. Her research interests include critical success factors of electronic commerce, information technology adoption, and consumer behavior in the digital market. Her prior research has been published in international journal and international conference such as ACM SIGecom Exchanges, International Journal of Electronic Customer Relationship Management, E-business and Telecommunications, and International Conference on Electronic Commerce.

Chotipong Saengkhunthod received his B.Sc. in Data Science and Business Analytics from the School of Information Technology, King Mongkut's Institute of Technology Ladkrabang. In 2021, he was hired as the System Analyst by True Digital Park and is currently the RPA Full-Stacked Developer at Metro Systems Corporation Public Company Limited.

Parischaya Kerdnoonwong received her B.Sc. in Data Science and Business Analytics from the School of Information Technology, King Mongkut's Institute of Technology Ladkrabang. In 2021, she was hired by NTT DATA (Thailand) Company Limited and is currently the Solution Consultant.

Hutchatai Chanlekha is an assistant professor at the Department of Computer Engineering, Faculty of Engineering, Kasetsart University, Thailand. She received her B.Eng and M.Eng degree in Computer Engineering from Kasetsart University. She received the Ph.D. in Informatics from the Graduate University for Advanced Studies, Japan. Her research interests include Natural Language Processing, applied Machine Learning, and Knowledge Engineering.

Nagul Cooharojananone is an associate professor at the Department of Mathematics and Computer Science, Faculty of Science, Chulalongkorn University, Thailand. He received his B.S. degree in Computer Science from Mahidol University. He received his M.Eng and Ph.D. in Information and Communication Engineering from the University of Tokyo. His research interests include Computer Vision, Machine Learning, Multimedia Technology and User Interface Design.