Detection of Unreliable Medical Articles on Thai Websites

Chotipong Saengkhunthod *, Parischaya Kerdnoonwong†, Kanokwan Atchariyachanvanich‡
Faculty of Information Technology,
King Mongkut's Institute of Technology Ladkrabang (KMITL), Bangkok, Thailand. 10520
E-mail: *60070133@it.kmitl.ac.th, †60070150@it.kmitl.ac.th, ‡kanokwan@it.kmitl.ac.th

Abstract—**Fake news have exerted terrible impact on the Thai society for a long time, especially fake health and medical news: unreliable news from social media have threatened people's mind and physical health. In this research, we investigated various methods for solving the problem of getting fake news on health and medical issues in social media. Then, we proposed to detect unreliable medical articles existed on Thai websites based on a machine learning. We collected samples of 297 reliable and 235 unreliable articles from 7 websites and analyzed the differences between them. Then, we selected 20 features that affected the reliability or unreliability of the articles and used machine learning to classify the articles according to those features. Experimental results show that XGBoost methods were the most effective at 90.60% accuracy.**

Keywords-Unreliable; Medical Articles; Machine Learning; text analytics

I. INTRODUCTION

Today, the exposure to pieces of information from social media and websites has changed because the competitive environment forces them to deliver news more rapidly, and more people are using the internet as the medium to get a variety of information.

From our exploration of the internet, we found that most pieces of fake news or false information were related to health and medical information. They scattered across all accessible social media and websites on the internet. A lot of people were persuaded into believing those pieces of false information from unreliable sources of information. Some of the false pieces of information were created to initiate a social trend that can induce the creation of a chain of consequential fake news. The deviation from scientific facts of these news was, sometimes, ignored, and the false information led to serious illness or even mortality.

In a way, this issue has been addressed. The Thailand's government has devised a way to solve it by creating an "Anti Fake News Center Thailand" [1] website. The website collected pieces of news and articles from sources on the Internet and analyzed them whether any of them are fake or not. However, the website only covered the issues in current trend at the time, so their analysis of fakeness did not span across all pieces of fake news and often not rapid enough to be useful.

Therefore, we attempted to use a new approach to tackle this issue. We modified and applied a data science procedure to more rapidly identify unreliable Thai medical articles. That is, when a user enters a suspected piece of fake news into their website and search for its fakeness, he or she may have to wait for an indefinitely long time, but we used the Machine Learning based framework to identify the fakeness. Our objectives were to investigate the article characteristics on Machine Learning techniques, then a data analysis model construction. The model was tested and the accuracy of model was showed in the confusion matrices.

This rest of this paper is arranged as follows: Section II describes related reference studies; Section III shows the detailed operational procedure of the Data Science Lifecycle [2] that we used; Section IV reveals the results of each model. The last section describes possible future projects and conclude the paper.

II. LITERATURE REVIEW

A. Detection of Fake News

Liu et al. used a machine learning technique to analyze and detect false medical information in China's social media [3], starting from collecting medical data from Chinese social media and categorizing them into 2 separate types: reliable and unreliable. Next, the article's title is analyzed—the words and their usage in the title can help identify whether an article is of the first or second category. Two methods were used for the analysis: a feature-based method and a content-based method. In the feature-based method, the researchers identify the features for the collected data such as the number of special characters and the writing styles. According to one of Liu et al.'s research works, 104 features were found and used by machine learning models such as SVM, k-NN, AdaBoost, GBDT, and Random forest. On the other hand, the content-based method is text-based. Researchers used FastText to analyze an article and used word sequence to differentiate the pattern and type of data of authentic and fake news. The best performance model using this method is GBDT, achieving 83.74% accuracy on test set.

In one of their research works, Ozbay and Alatas used a supervised Machine Learning technique to predict false information on social media. That work consisted of three steps [4]. Firstly, they collected articles with false information on social media. Secondly, they pre-processed the data by using a Word tokenization method to remove unnecessary words and then created the term frequency and document term matrix. Lastly, they evaluated the model and reported the most efficient algorithm for that purpose which was Decision Tree, with 74.5% mean accuracy.

Songram used a Machine Learning technique to detect unreliable Facebook pages [5]. That work comprises 3 steps. Firstly, Naive Bayes, k-NN and SVM were selected as classification models. Secondly, features selection was performed with 4 methods: Gini index, Chi2, Fisher, and Lasso. Finally, the model was evaluated with a 10-fold cross validation and reported that the best performing feature selection method was Fisher, with 91.37% accuracy.

978-1-7281-7602-4/21/$31.00 ©2021 IEEE

B. Health Information Websites

In one of their research papers, Samuel and Zaiane explored about attributes of reliable medical websites [6]. Those attributes were grouped into two categories: article-based and community-based. Information provided by article-based websites was written by experts, and each article can be validated by cross-checking with a confirmed valid article. Information provided by community-based websites can be accessed by anyone. Members of this kind of websites can vote, give score, and comments on the validity of an article.

III. MEDTHODOLOGY

We conducted our research according to a data science lifecycle method [2] as follows.

A. Business Understanding of the information

According to the method, we needed to understand the problems of false medical information clearly in terms of business understanding. Our understanding covered the effects of receiving and distributing false information. Then, we needed to formulate a solution by generating a machine learning model that can detect pieces of reliable information from unreliable ones to analyze false information for users.

B. Data Acquisition and Understanding them

1) Reliable media article: According to a paper by Samuel et al. [6] that reports about choosing the right sources of articles—a reliable article must be written by experts and verifiable. We used this approach to select our sources of reliable information from official websites of Sukhumvit hospital, Ramathibodi hospital, and Medical Faculty of Mahidol University, Thailand. The total number of authentic articles was 297.

2) Unreliable medical articles: We used Zhang and Ghorbini's approach [7] in collecting unreliable information from websites that were not moderated by experts and articles from unreliable sources such as dokkaew.wordpress.com, bamgpunsara.com, colonocancerzone.com, and eatonlinehealth.com and considered a total of 235 unreliable articles.

We used a Python library called "Beautiful soup" and a web scraping technique to collect the data that we have already searched for from various websites and select a total of 7 websites and 532 reliable and unreliable articles. Details of our dataset are shown in Table I.

C. Data Cleaning

We found a lot of unwanted elements in our data after doing a web scraping technique. To prevent modeling errors, those unwanted elements had to be removed. Articles with those elements were such as comments on "Medthai" and articles with unwanted "\n" or "\t" from "Bangpunsara" and "Dokkaew".

TABLE I. RELIABLE AND UNRELIABLE SOURCES OF INFORMATION

Site Name	URL	Number of articles
Medthai	https://www.medthai.com	118
Sukumvit Hospital	http://www.sukumvithospital.com	88
Med Mahidol	https://med.mahidol.ac.th	91
Total of Reliable articles		297
Dokkaew	https://dokkaew.wordpress.com	56
Bangpunsara	http://www.bangpunsara.com	53
Coloncancerzone	https://www.coloncancerzone.com	88
Eatonlinehealth	http://www.eatonlinehealth.com	38
Total of Unreliable articles		235

D. Data Cleaning

We found a lot of unwanted elements in our data after doing a web scraping technique. To prevent modeling errors, those unwanted elements had to be removed. Articles with those elements were such as comments on "Medthai" and articles with unwanted "\n" or "\t" from "Bangpunsara" and "Dokkaew".

E. Modeling

We developed an unreliability prediction model by modifying the framework for unreliable medical article analysis proposed by Liu at el. [3].

Fig. 1 shows the components of framework for unreliable medical article analysis: Feature Engineering, Model training, and Model evaluation.

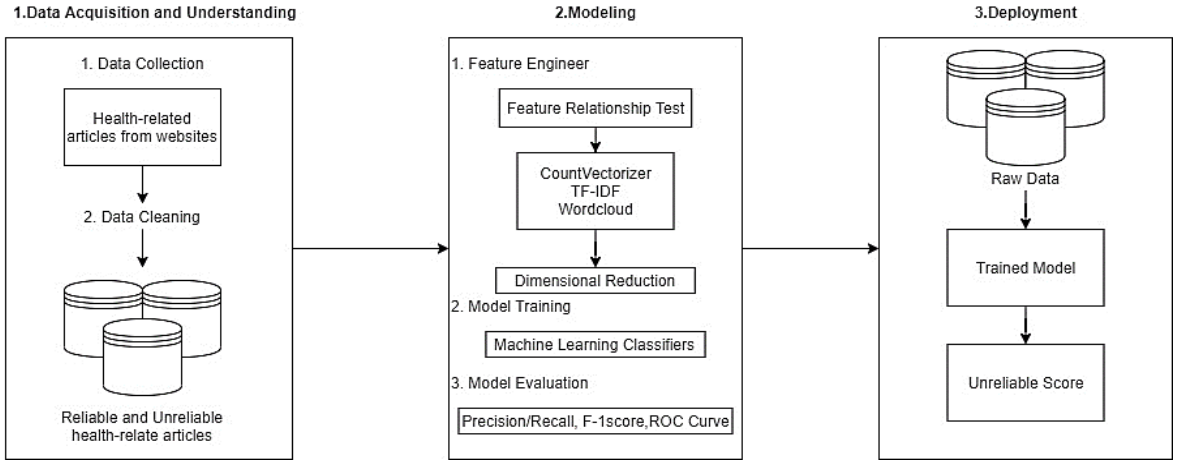

Fig. 1. Framework for Unreliable Medical Article Analysis

1) Feature relationship:

- *Feature relationship test:*

We investigated required features for classifying reliable and unreliable articles. We selected a few features from Liu et al.'s paper [3] and used them to do a Chi-square test to determine if there is any relationship between the selected features and the reliability of the article. However, a lot of data for some features had values of 0, so we used a Dummy method to transform them into values of 1, and if the data was higher than 0 in the first place, we used an Information Gain method to evaluate the impact level of each of those features, as illustrated in Fig. 2

```
Feature : num_of_pic is in relation(IG = 0.1974)
Feature : number_'!'_title is in relation(IG = 0.0816)
Feature : number_'?'_title is in relation(IG = 0.0166)
Feature : number_':'_title is in relation(IG = 0.0052)
Feature : amount_num_title is in relation(IG = 0.0252)
Feature : number_'!'_text is in relation(IG = 0.0184)
Feature : number_'?'_text is in relation(IG = 0.0462)
Feature : number_'[]'_text is in relation(IG = 0.0812)
Feature : text_length is in relation(IG = 0.6602)
Feature : pic_per_text is in relation(IG = 0.6108)
Feature : amount_num_text is in relation(IG = 0.1977)
Feature : num_of_ref is in relation(IG = 0.3402)
```

Fig. 2. Information Gain method for evaluating the impact level of each selected feature

- *Natural Language Processing (NLP):*

We used NLP to cull out words from an article and put them through Bag-of-Words (BoW) and Term Frequency Index Frequency (TF-IDF) processes to find unreliability-indicating words in that article. Additionally, we used TF-IDF to create a word cloud that can detect words that can demonstrate the most frequent words found in the reliable articles. Some reliability-indicating words detected by word cloud were "มะเร็ง (cancer)", "อาการ (symptom)", "คนป่วย (patient)", and "โรค (disease)". More of them are shown in Fig. 3 The title of a reliable article often includes the following words: "วิธี (method)", "อาการ (symptom)", "สาเหตุ (cause)", "การรักษา (treatment)", and "โรค (disease)". More of them are shown in Fig. 4

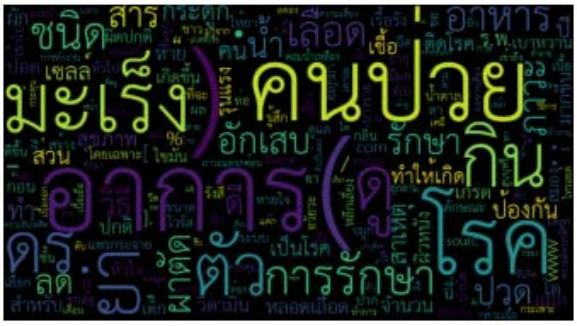

Fig. 3. Reliability-indicating words outputted by word cloud

Fig. 4. Reliability-indicating words in the title of reliable articles outputted by word cloud

Unreliability-indicating words detected by word cloud were such as "อาการ (symptom)", "มะเร็ง (cancer)", "สมุนไพร (herb)", and "ปวด (pain)". More are shown in Fig. 5 Unreliability-indicating words in the title of unreliable articles detected by word cloud were such as "อาการ (symptom)", "มะเร็ง (cancer)", "อาหาร (food)", and "ลด (reduce)". More are shown in Fig. 6

Fig. 5. Unreliability-indicating words outputted by word cloud

Fig. 6. Unreliability-indicating words in the title of unreliable articles outputted by word cloud

word cloud also pointed to significant words in the content of articles for classifying reliable and unreliable articles, including "ฝังตัว (implant)", "น้ำเกลือ (saline)", and "จำนวน (amount)". More are shown in Fig. 7 As well, it also pointed to significant words in the title of articles for classifying reliable and unreliable articles such as "หัด (measles)", "ฝุ่น (dust)", and "ผอม (thin)". More are shown in Fig. 8

Fig. 7. Word cloud of significant words in the content of articles for classifying reliable and unreliable articles

Fig. 8. Word cloud of significant words in the title of articles for classifying reliable and unreliable articles

The results from word cloud led us to select more features affecting the reliability and unreliability of the articles, resulting in better identification of reliable data from unreliable data. We added new words, "มะเร็ง (cancer)" and "อาการ (symptom)", outputted from the BoW process, and "ผอม (thin)", "ประโยชน์ (useful)", "วิธีใช้ (instruction)", "ควรจะ (should)", "สกัด (extract)" and "&" outputted from the TF-IDF process into a group of new features to be analyzed. We added these new features to our model, tested it, and obtained the results shown in Fig. 9.

```
Feature : number_'ผอม'_title Can use by info_gain method IG = 0.0019
Feature : number_'ประโยชน์'_title Can use by info_gain method IG = 0.0053
Feature : number_'วิธีใช้'_title Can use by info_gain method IG = 0.0019
Feature : number_'ควรจะ'_title Can use by info_gain method IG = 0.0019
Feature : number_'สกัด'_title Can use by info_gain method IG = 0.0031
Feature : number_'&'_title Can use by info_gain method IG = 0.0126
Feature : number_'มะเร็ง'_title Can use by info_gain method IG = 0.1427
Feature : number_'อาการ'_title Can use by info_gain method IG = 0.0130
```

Fig. 9. New features output from TF-IDF

Only feature scattered in most of reliable and unreliable articles and its information gain is higher than 0 will be selected as feature. Since the information gain of these features: "ฝังตัว (embed)", "จำนวน (number)", หัด (measles)", "ฝุ่น (dust)" are 0.00, 0.00, 0.00008, and 0.0003 respectively, they were not added as feature set. A total of 20 features, listed in Table II, were used in the completely enhanced model.

TABLE II. FEATURE SET

Feature Set	
Word frequencies	"มะเร็ง (cancer)", "อาการ (symptom)", "ผอม (thin)", "ประโยชน์ (useful)", "วิธีใช้ (instruction)", "ควรจะ (should)", "สกัด (extract)"
Others	"!", "?", ":", "&" in title "!", "?", "[ ]" in text Amount of numbers in title/text Text length of each paragraph Number of picture in paragraph Number of characters per picture in text Number of reference in paragraph

- *Dimensionality reduction:*

We reduced the dimensions of the data to two dimensions by using a t-distributed Stochastic Neighbor Embedding (t-SNE) technique that showed the grouping of the data clearly. Therefore, we could see the distribution of the data and analyze them (see Fig. 10).

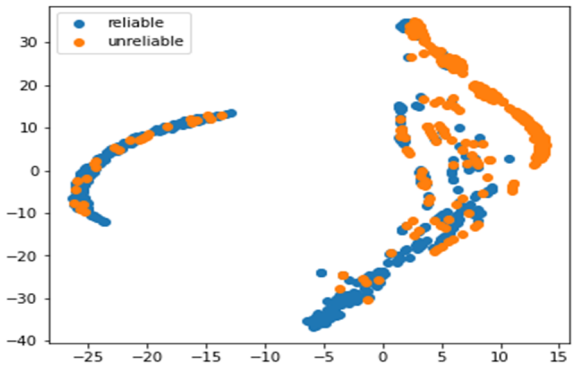

Fig. 10. Grouping of the data after dimensionality reduction

2) Model Training:

We used different machine learning models from those used by Liu at el. [3]. We selected XGBoost algorithm over AdaBoost and GBDT because XGBoost could handle an overfitting problem better than AdaBoost and GBDT [8]. We did not select Random Forest because this algorithm creates many complex trees which are difficult to visualize and understand. On the other hand, objects of Decision Tree were easier to visualize and understand [9]. We selected Logistic Regression because the data contained a lot of numerical values.

We split 70 percent of the whole data into a training set and 30 percent into a testing set. We used five classification models: Logistic Regression, Support Vector machine with linear kernel and C value equal to 0.1, k-Nearest Neighbors with k-value equal to 3 by using elbow method to define the k-value, Decision Tree and XGboost.

- *Support Vector Machine (SVM):*

The main function of a support vector machine algorithm is to find a hyperplane in an N-dimensional space (N — the number of features) that distinctly classifies the data points. [10] Many hyperplanes for separating two classes of data points are possible. The best one that completely and distinctly separate the data points should be selected. Moreover, it needs to be a hyperplane with the maximum margin, i.e., maximum distances between data

points of both classes. Maximizing the margin distance provided some reinforcement so that future data points could be classified with more confidence.

- *Decision Tree:*

A decision tree is a decision support tool using a tree-based model or if-then rule to generate prediction. [11] It can be used to visually and explicitly represent decisions and decision makings that contains conditional control statements.

- *XGBoost:*

XGBoost is a tree-based model by optimized Gradient Boosting. [12] XGBoost provides a parallel tree boosting with more accurate approximated values by employing second-order gradients and advanced regularization.

- *k-Nearest Neighbor (k-NN):*

k-NN is a type of instance-based learning, or lazy learning, it is a majority vote of its neighbors with measured by distance function. [13] The input consists of the k-closest training examples in the feature space. The output depends on whether k-NN is used for classification or regression.

- *Logistic Regression:*

Logistic regression is a statistical model that, in its basic form, uses a logistic function to model a binary dependent variable. [14] The output of the hypothesis is the estimated probability. This is used to infer how confident can predict value be actual value when given an input. Mathematically, a binary logistic model has a dependent variable with two possible values, such as pass/fail which is represented by an indicator variable, where the two values are labeled "0" and "1".

3) Model Training:

We held Ward's research [15] as a reference for evaluating the performance of each model to help us determine the best model for classification in terms of Precision, Recall, F1 score and ROC curve.

- *Precision:*

Precision is the fraction of retrieved documents that are relevant to the query,

$$Precision = \frac{TP}{TP + FP} \quad . \tag{1}$$

- *Recall:*

Recall is the fraction of the relevant documents that are successfully retrieved,

$$Recall = \frac{TP}{TP + FN} \quad . \tag{2}$$

- *F1 score:*

F1 score is the harmonic mean of precision and recall,

$$F_1\,score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \quad . \tag{3}$$

- *Receiver operating characteristic Curve (ROC):*

ROC curve is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. An ROC curve approaching 1 signifies a high performance of the model.

4) Deployment:

Business users could use the outcomes of the Analysis of Unreliable Medical Articles on Thai Websites to fulfill their present and future needs. They could query about an article on the website and get the probability that the article was unreliable.

IV. Evalution Results

From performance evaluation of all models, we created Table III. It contains the result for each model, sorted according to the accuracy value from high to low.

TABLE III. Results From Each Model

Model	Accuracy	Precision	Recall	F1 score
XGBoost	90.60%	90.98	89.05	89.83
Decision tree	88.59%	88.02	87.76	87.88
SVM	86.58%	85.68	86.13	85.88
Logistic regression	83.22%	82.45	81.74	82.05
k-NN	81.21%	80.07	80.78	80.36

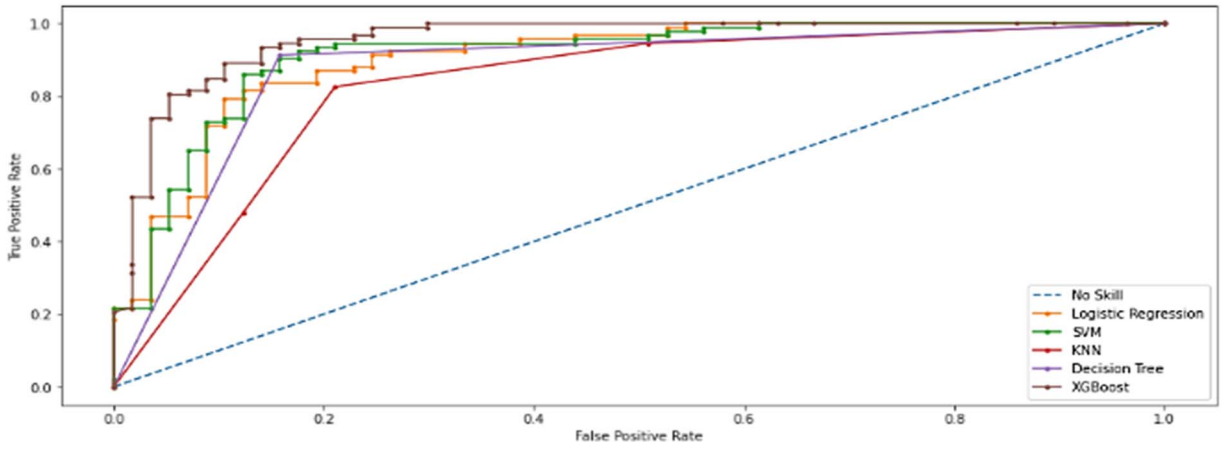

Fig. 11. ROC Curve results

As can be seen in Fig. 11, the ROC curve of XGBoost approaches 1 more closely than that of the Decision tree does. The closer the curve is to 1, the higher the model's performance is.

From the experiment with each model, we saw that the selected features were able to differentiate reliable and unreliable data from each other quite accurately. The most accurate models were XGBoost with 90.60% accuracy.

V. Conclusion and Future Work

In this research, we detected unreliable medical articles on Thai websites by analyzing writing style, word usage, frequency of words and special characters. We selected the features referred to by Liu at el. [3]. We used Chi-square and Information gain to test the relationship between features and target variable. In addition, we created word clouds from BoW and TF-IDF to observe words that could indicate the reliability or unreliability of an article. Words that we obtained by using those new features are such as "มะเร็ง (cancer)", "อาการ (symptom)", "ผอม (thin)", "ประโยชน์ (useful)", "วิธีใช้ (instruction)", "ควรจะ (should)" and "สกัด (extract)"

We used classification models including Logistic Regression, Support Vector machine, k-Nearest Neighbors, Decision Tree, and XGBoost to classify the reliability or unreliability of medical articles. The classification output indicated that XGBoost were the most accurate classification models.

Possible future works includes the following: we will collect more medical articles to train our model and predict future data; we will add more features to improve the classification accuracy; and we will develop the model further to be ready for real-life usage.

References

[1] "Anti-Fake News Center Thailand." https://www.antifakenewscenter.com/ (accessed Oct. 21, 2020).

[2] S. Siva, "Data Science life Cycle | Towards Data Science." https://towardsdatascience.com/stoend-to-end-data science-life-cycle-6387523b5afc (accessed Oct. 21, 2020).

[3] Y. Liu, K. Yu, X. Wu, L. Qing, and Y. Peng, "Analysis and detection of health-related misinformation on Chinese social media," IEEE Access, vol. 7, pp. 154480–15448019, doi: 10.1109/ACCESS.2019.2946624.

[4] F. A. Ozbay and B. Alatas, "Fake news detection within online social media using supervised artificial intelligence algorithms," *Phys. A Stat. Mech. its Appl.*, vol. 540, p. 123174, 2020, doi: 10.1016/j.physa.2019.123174.

[5] P. Songram, "Detection of unreliable and reliable pages on Facebook," *Artif. Life Robot.*, vol. 24, no. 2, pp. 278–284, 2019, doi: 10.1007/s10015-018-0509-z.

[6] H. W. Samuel and O. R. Zaiane, "PSST... privacy, safety, security, and trust in health information websites," *Proc. - IEEE-EMBS Int. Conf. Biomed. Heal. Informatics Glob. Gd. Chall. Heal. Informatics, BHI 2012*, pp. 584–587, 2012, doi: 10.1109/BHI.2012.6211650.

[7] X. Zhang and A. A. Ghorbani, "An overview of online fake news: Characterization, detection, and discussion," *Inf. Process. Manag.*, vol. 57, no. 2, p. 102025, 2020, doi: 10.1016/j.ipm.2019.03.004.

[8] SauceCat, "Boosting algorithm: XGBoost. This article continues the previous… | Towards Data Science." https://towardsdatascience.com/boosting-algorithm-xgboost-4d9ec0207d (accessed Nov. 02, 2020).

[9] A. Navada, A. N. Ansari, S. Patil, and B. A. Sonkamble, "Overview of use of decision tree algorithms in machine learning," *Proc. - 2011 IEEE Control Syst. Grad. Res. Colloquium, ICSGRC 2011*, pp. 37–42, 2011, doi: 10.1109/ICSGRC.2011.5991826.

[10] R. Gandhi, "Support Vector Machine — Introduction to Machine Learning Algorithms." https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47 (accessed Oct. 23, 2020).

[11] "Decision Tree - Overview, Decision Types, Applications." https://corporatefinanceinstitute.com/resources/knowledge/other/decision-tree/ (accessed Oct. 23, 2020).

[12] D. Ananda, "How does XGBoost Work. Understanding the internal working of XGBoost" https://towardsdatascience.com/how-does-xgboost-work-748bc75c58aa (accessed Oct. 23, 2020).

[13] T. Srivastava, "K Nearest Neighbor | KNN Algorithm | KNN in Python & R." https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/ (accessed Oct. 23, 2020).

[14] S. Gupta, "What makes Logistic Regression a Classification Algorithm?" https://towardsdatascience.com/what-makes-logistic-regression-a-classification-algorithm-35018497b63f (accessed Oct. 23, 2020).

[15] D. M. W. Powers and Ailab, "Evaluation : From Precision , Recall and F-Factor to ROC , Informedness , Markedness & Correlation," *J. Mach. Learn. Technol.*, vol. 2, pp. 37–63, 2007, doi: 10.9735/2229-3981.