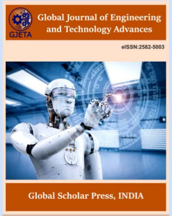
GJETA

Global Journal of Engineering and Technology Advances

eISSN: 2582-5003

Cross Ref DOI: 10.30574/gjeta

Journal homepage: https://gjeta.com/


Global Journal of Engineering and Technology Advances
GJETA
eISSN:2582-5003
Global Scholar Press, INDIA

(Research Article)

Check for updates

Forecasting onion armyworm using tree-based machine learning models

Marcelino Concepcion Collado Jr [1, 2, *] and Gilbert Malawit Tumibay [2]

[1] *College of Information and Communications Technology, Nueva Ecija University of Science and Technology, Cabanatuan City, Nueva Ecija, Philippines.*
[2] *Graduate School, Angeles University Foundation, Angeles City, Philippines.*

Global Journal of Engineering and Technology Advances, 2023, 15(03), 001–007

Publication history: Received on 17 April 2023; revised on 30 May 2023; accepted on 02 June 2023

Article DOI: https://doi.org/10.30574/gjeta.2023.15.3.0095

Abstract

In the Philippines, the province of Nueva Ecija produces fifty-four percent of its annual onion production. However, the level of onion growth production was reduced; since the outbreak of 2016, armyworms destroyed thousands of hectares of farms resulting in a loss of billions of pesos, which lead to the decline of the onion harvest. In this study, we develop machine learning models to forecast an outbreak of armyworms to help evade or reduce the damage caused by an armyworm outbreak. Climatic data; particularly Maximum temperature, Minimum Temperature, Ultraviolet Index, Humidity, Cloudiness, Wind Speed, Sun Hours, Rainfall, and Pressure from the Philippine Atmospheric, Geophysical and Astronomical Services Administration (PAGASA) and armyworm outbreak occurrences data from the Provincial Agriculture Office (PAO) of Nueva Ecija was used as the dataset for this study Using Tree-based machine learning models Decision Tree and Random Forest. Binary classifiers were developed and evaluated to forecast the occurrence or non-occurrence of the armyworm outbreak and the use of feature importance to distinguish the most critical climatic features that significantly contribute to forecasting an armyworm outbreak in the province of Nueva Ecija. These tree-based models produced satisfactory results, with the Random Forest model exhibiting a better forecasting capability than the Decision Tree model.

Keywords: Machine Learning (ML); Algorithms; Tree-Based Model; Decision Tree Model; Random Forest Model; Forecasting; Armyworm

1. Introduction

In the Philippines, climate change has become a significant threat to the agricultural industry. In 2016 an outbreak of Armyworm *Spodoptera exigua* (Hübner) (Lepidoptera: Noctuidae) occurred in the onion farms in the province of Nueva Ecija. As triggered by the El Niño phenomenon, long-distance migration from other countries northeast of the Philippines may have caused the pest outbreak [1]. The province of Nueva Ecija generates fifty-four percent of the country's onion annual output [2]. Armyworm appears in the various farms giving substantial harm to onions, which leads to increased expense in farm production and crop loss [3]. The reduction of crops caused by armyworms is a significant challenge for the country's efficient and sustainable food production.

To control the armyworm population, farmers spray fields with pesticides before the pest create a considerable loss on the onion crops. However, factors that include the use of illegal insecticides and pesticides' mishandling contribute to more damage [3]. As stated by onion farmers, the pests are immune to chemical spray, implying pesticide resistance development due to malpractices in pesticide application [1]. The time when to spray pesticides plays a crucial role in the pesticide's effectiveness. For this purpose, a careful prediction is vital to improve the timing of pest control measures to avoid crop losses and the excessive use of pesticides.

[*] Corresponding author: Marcelino Concepcion Collado Jr

Climate influences the proliferation of agricultural pests. Since temperature, light, and water are major factors that regulate their growth and development [4]. Since these armyworms are weather-dependent, a climate-based forecasting model can help develop pest management intervention plans. When a prediction of armyworm infestation occurs, farmers can make early pest management measures to protect the onion crops and minimize losses in advance. Forecasting a crop pest at an early stage using machine learning technology would be beneficial.

Machine learning is defined as the science of programming computers to learn from data; it implements the scientific basis of data mining. A subset of artificial intelligence, machine learning can interpret the acquisition of structural descriptions from patterns. The kind of descriptions found can be used for prediction, explanation, and understanding. Applications focused on prediction; can forecast what will happen in new situations from data that describe what happened in the past, often by predicting the classification of new examples [5].

The growing popularity of machine learning applied in agriculture has increased. This study presents a forecasting solution that involves tree-based machine learning models that rely on climate data to forecast armyworm occurrences. The researchers considered using tree-based machine learning models to analyze the climatic factors that influence the armyworm outbreak. For the past years, there is more emphasis on research studies on the relevance of machine learning techniques in crop pest prediction. Machine learning implemented in the agricultural sector can improve conventional pest management techniques in the most cost-friendly approach.

In this study, we analyze the climate factors that correlate with an onion armyworm outbreak occurrence in the province of Nueva Ecija in the Philippines. Tree-based model classifiers were built to forecast whether an armyworm outbreak will occur during a given weather situation. Finally, we evaluate and compare the predicted accuracies of the tree-based machine learning models.

1.1. Related Works

The outcome of Arulkumar et al.'s research proved that weather conditions have a significant influence on the beet armyworm population and migration. Onion is a major crop in India. However, the arrival of beet armyworm affected the country's onion harvest. Their results have shown a correlation between the damages caused by beet armyworm and the climatic patterns. The result of their study can aid in scheduling management plans in onion produce against beet armyworm. In constructing a weather-based pest forewarning system, data regarding prevailing weather patterns are essential[6].

Calderon et al. utilized data mining methods in the agriculture industry to predict the rice black bug epidemic using a decision tree algorithm. Their research dataset comprises historical rice black bug infestation occurrences from different municipalities and cities in Quezon, Philippines, from 2005 to 2015, and climatic data such as temperature and humidity, lunar cycle, and soil status[7]. Combining the methods of Insect Epidemiology Data Mining in agriculture and decision tree algorithm enabled them to predict pest occurrence outcomes. Their study proves that data mining in agriculture and tree-based models, such as a decision tree algorithm, can help create predictive models for pest occurrence.

Balaban et al. developed a machine learning solution that can replace conventional methods for determining the Sunn Pest pesticide application time. By training the machine learning models on four years of weather data and Sunn Pest occurrence data. Farmers can decide if the fields must be sprayed with insecticide by pairing climatic data with the Sunn Pest phases in their life cycle. The Sunn pest is a bug found in the Middle East region that causes significant harm to wheat. Their solution uses two classes of tree-based machine learning models. The first utilizes a decision tree algorithm for predicting the Sunn Pest migration pattern from winter hives to wheat fields. The second utilizes a random forest algorithm that predicts Sunn Pest's sexual maturity period, which is a deciding factor for applying pesticides. Using the voting mechanism between the Random forest trees minimized overfitting in their predictive model[8]. The study showed that using a tree-based algorithm to construct a prediction model could generate accurate predictions.

2. Material and method

2.1. Data Acquisition

The Provincial Agriculture Office (PAO) of Nueva Ecija provided the data regarding the incidence of armyworm outbreak for this study. Since 2016, PAO conducted field monitoring in different municipalities of Nueva Ecija to observe armyworms' existence. The PAO aims to develop methods to prevent the outbreak of armyworms in the province. For each municipality, data were given based on the incidence of the armyworm outbreak each year.

Historical climate data from the Philippine Atmospheric, Geophysical and Astronomical Services Administration (PAGASA) and Agrometeorological (Agromet) weather stations have been obtained from the following locations; Cabanatuan City PAGASA Station and Munoz Nueva Ecija Agromet Station. The historical climate data spanned from 2016 to 2019 and contained the following attributes: Maximum temperature, Minimum Temperature, Ultraviolet Index, Humidity, Cloudiness, Wind Speed, Sun Hours, Rainfall, and Pressure.

2.2. Data Preprocessing

The data collected from OPA were appended to the PAGASA data, combined into a single dataset. Each row of OPA data documenting a specific Armyworm outbreak report will correspond with the PAGASA climate data. Additional rows are also included in the dataset to reflect periods when the armyworm outbreak did not occur. The merged table will serve as the research dataset, which was subjected to data preprocessing. The missing values in the dataset were handled using the Scikit learn preprocessing library. Data preprocessing is a required step in any machine learning research because the raw data might have missing values or many discrepancies in each feature, leading to erroneous results[9]. One of the many characteristics of the decision tree algorithm is that they need very little data preparation; in particular, since they are not sensitive to the data's variance, they do not require feature scaling[10]. The dataset has 750 instances, including 262 rows of the armyworm outbreak and 488 rows of Non-occurrence of the armyworm outbreak. In the Target attribute of the dataset, Non-occurrence was replaced by 0 and occurrence by 1. The preprocessed data that compiled the climatic factors and armyworm outbreak dataset were divided into training and testing datasets. This step includes splitting the series into training and testing of the data set. The dataset has a slight imbalance; therefore, we used Synthetic Minority Oversampling Technique (SMOTE) [11] for resampling to balance the training dataset.

2.3. Tree-Based Machine Learning Models

Tree-based machine learning models were utilized to predict the occurrence or non-occurrence of armyworm infestation. The machine learning models built in this study will use state-of-the-art classification machine learning algorithms under supervised learning, Decision Tree classifier, and Random Forest classifier, which are briefly described in the following sections.

2.3.1. Decision Tree Classifier

Classification and Regression Trees (CART) refer to Decision Tree algorithms used for predictive modeling problems of classification or regression. The Decision Tree algorithm provides the basis for essential algorithms such as bagged decision trees, random forests, and boosted decision trees[10]. Decision Tree aims to divide the dataset into subsets by attribute by the instance with the closest relationship with the target variable. The process is repeated until the specified rule is fulfilled convincingly. The Decision Tree algorithm is considered the most frequently used algorithm for identifying crop pests and diseases and the best to produce rapidly interpretable classifiers[12].

2.3.2. Random Forest Classifier

Random Forest is an ensemble of decision trees that takes on the concept of bagging, which is short for bootstrap aggregation, where each model in the ensemble is used to generate a prediction for a new sample [13]. Bagging is used to reduce the variance of an estimated prediction function[14]. In Random Forest, each tree casts a vote for the classification of a new sample so that the predictions of each tree are averaged to give the forest's prediction[15]. Random forest models have resulted in significant improvements in prediction accuracy compared to a single decision tree by growing ' n' number of trees; each tree in the training set is sampled randomly without replacement [13].

An evaluation of each climate feature's relative importance can be made using the built-in feature importance attribute of Random Forest by looking at how often the tree nodes use that feature to decrease impurity on average.

3. Results and Discussion

This section presents the results obtained from the Decision Tree and Random Forest classifiers. All experiment processes are performed with Python 3.7 on a Laptop with 1.10GHz Intel Pentium Silver N5000 CPU and 4GB RAM, running Windows 10 Operating System.

3.1. Evaluation Metric for Model Performance

The model's performance is evaluated using the following metrics; Confusion matrix, Accuracy, Precision, Recall, and F1-score. The confusion matrix will serve as an evaluation tool for analyzing how well the classifiers can recognize if the model is confusing two classes by presenting the correct and incorrect predictions than the actual classifications in the test data [16]. Accuracy will be measured for the tree-based models to determine the occurrence or non-occurrence of the armyworm outbreak; below is the equation for accuracy.

$$Acc = (TP + TN)/(TP + FP + FN + TN)$$

Where:
TP = The number of True Positives
TN = The number of True Negatives
FP = The number of False Positives
FN= The number of False Negatives

Precision measures the proportion of predicted positive cases of armyworm outbreak that are correctly predicted; below shows the equation for precision.

$$\text{Precision} = \text{TP}/(\text{TP}+\text{FP})$$

Recall measures the proportion of positive cases of armyworm outbreak correctly predicted below shows the equation for recall.

$$\text{Recall} = \text{TP} / (\text{TP} + \text{FN})$$

The F1-score is the harmonic mean of precision and the recall; below shows the equation for the F1-score.

$$\text{F1-score} = 2*(\text{ Precision} * \text{Recall})/(\text{ Precision} + \text{Recall})$$

3.2. Feature Importance

Using Random Forest to extract the feature importance, we can explore the most significant climate features that affect the Random Forest model in classifying its forecast as an occurrence or non-occurrence of an armyworm outbreak. We may visualize and measure the importance to see which features are most significant to the model. Figure 1 shows the following feature importance.

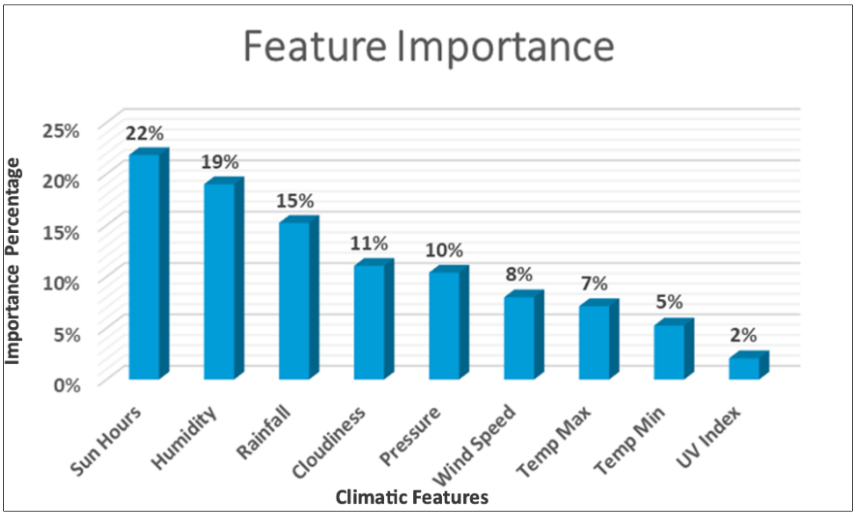


Figure 1 Feature Importance

Sun Hours, which measures sunshine duration in a month, accounts for 22 percent importance in the model's prediction of the armyworm outbreak's occurrence. Followed by humidity, which is the monthly average amount of vapor in the air, accounts for 19 percent. Rainfall, which is the average amount of showery precipitation for the month, is 15 percent. Cloudiness, which is the average amount of clouds in a month, is 11 percent. The pressure, which is the atmospheric pressure, had a value of 10 percent; pressure is an indicator of the weather; when an area has low-pressure, it leads to cloudiness, wind, and precipitation. Usually, high-pressure contribute to fair weather. Wind Speed, which is the average airspeed per month, had 8 percent in the feature importance. The maximum temperature is 7 percent, while the minimum temperature is 5 percent; both are measured in Celsius. Ultraviolet (UV) Index, or the average radiant energy released by the sun, has the lowest feature importance with a score of 2 percent.

Feature importance provides a better insight into the Random Forest model. Analyzing the dataset shows that higher armyworm outbreak incidents happened during periods of long sunny days. The Philippines' climate can be divided into two major seasons the rainy season and the dry season. As stated by the OPA, Weather patterns, characterized by a hot spell with brief intervals of rain, have become conducive to the development of pests that feed on onions[2].

3.3. Decision Tree Model

We used a Decision Tree classifier with the following parameters when constructing our first tree-based model; the minimum number of samples needed to break an internal node. The function for measuring the quality of a split. The maximum tree depth. The strategy used at each node to select the split. The minimum number of samples must be at a leaf node and the number of features to consider when looking for the best split. The test outcome of our model Decision Tree is shown in Table 1. This demonstrates that the model has an 88 percent forecasting accuracy, an 82 percent precision score, an 83 percent recall score, and an F1 score of 83 percent.

Table 1 Decision Tree Model

Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Decision Tree	88%	82%	83%	83%

A total of 188 samples were used as a test set for the study, 122 samples are non-occurrence of armyworm outbreak, and 66 samples are for armyworm outbreak occurrences. In the Decision Tree confusion matrix, the model was able to classify 110 correct non-occurrences, incorrectly classified 12 instances of non-occurrence as an armyworm outbreak. Furthermore, 11 armyworm instances were incorrectly classified as non-occurrence, while the model correctly classified 55 instances of armyworm outbreak. The result's Confusion matrix is shown in figure 2, where 0 stands for Non-occurrence of the armyworm outbreak and 1 for armyworm outbreak occurrences.

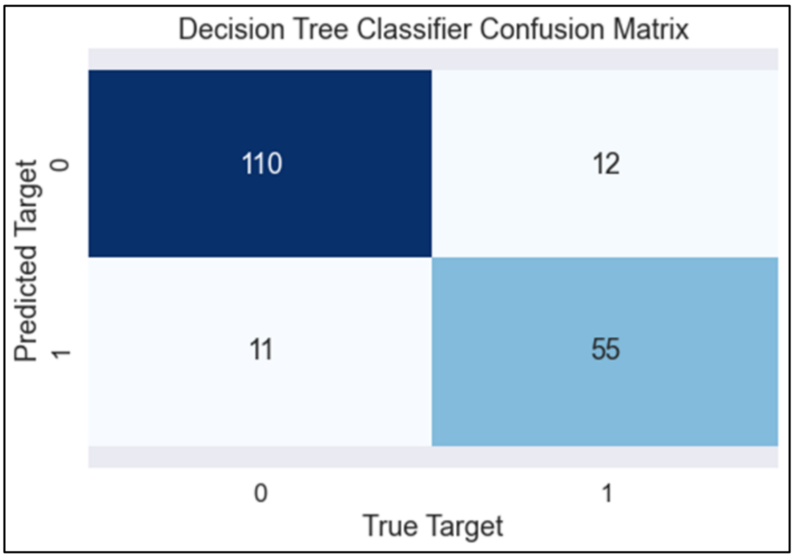


Figure 2 Decision Tree Confusion Matrix

3.4. Random Forest Model

Building our second tree-based model, we used an ensemble of Decision Tree classifiers known as Random Forest. The following parameters were used for this model; The number of trees in the forest. The minimum number of samples required to split an internal node. The minimum number of samples required to be at a leaf node. The tree's maximum depth and the number of features to consider when looking for the best split. The following shows results obtained using the Random Forest classifier on the dataset. Table 2 shows the test result of our Random Forest model. It shows that the model has an 89 percent forecasting accuracy, an 84 percent precision score, an 85 percent Recall score, and an 84 percent achieved F1-score.

Table 2 Random Forest Model

Model	Accuracy	Precision	Recall	F1-Score
Random Forest	89%	84%	85%	84%

A total of 188 samples were used as a test set for the study, 122 samples are non-occurrence of armyworm outbreak, and 66 samples are for armyworm outbreak occurrences. The Random Forest model was able to classify 111 correct non-occurrences, while the model incorrectly classified 11 instances of non-occurrence as an armyworm outbreak. Furthermore, 10 armyworm outbreak samples were incorrectly classified as non-occurrence, while the model correctly classified 56 instances of armyworm outbreak. Figure 3 shows the Random Forest confusion matrix where 0 stands for Non-occurrence of the armyworm outbreak and 1 for armyworm outbreak occurrences.

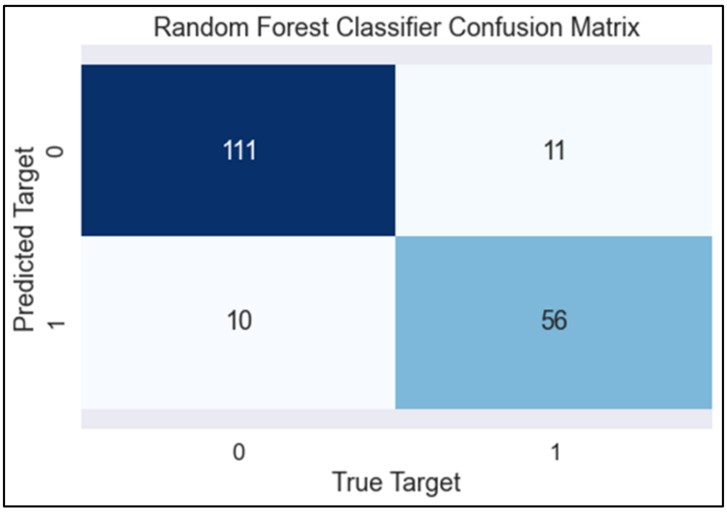


Figure 3 Random Forest Confusion Matrix

4. Conclusion

This paper presents a machine learning study on forecasting the armyworm outbreak in the province of Nueva Ecija, Philippines, by utilizing climatic data and the armyworm outbreak data, using tree-based machine learning models to predict the occurrence of armyworms. Our study's findings showed that the Random Forest Model achieved better overall forecasting performance than a single Decision Tree Model. Feature Importance showed that the top five climate features that affect the model's prediction of armyworm occurrence are Sun hours, Humidity, Rainfall, Cloudiness, and Pressure. This information can now help the decision-makers, agricultural organizations, and farmers provide better interventions in the control and management of the armyworm outbreak. However, there are still drawbacks to the proposed model, as the tree-based models were made based on a dataset with a small number of occurrences and features that may influence the models forecasting capability. Such an issue may prevent the proposed model from producing better performance. For the researchers to do a more comprehensive experiment for further developing the models, the armyworm outbreak dataset should be expanded to improve the models' predictive capability. Therefore, a continuous collection of armyworm data would push the project to a more robust forecasting system.

Compliance with ethical standards

Acknowledgments

The authors would like to express their gratitude and thanks to the Office of the Provincial Agriculturists (OPA) in the province of Nueva Ecija, Philippines.

Disclosure of conflict of interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Navasero MM, Cayabyab B, Ebuenga MD, Candano RN, Bautista NM, et al. Investigation on the 2016 outbreak of the onion armyworm, Spodoptera exigua (Hübner) (Lepidoptera: Noctuidae), in onion growing areas in Nueva Ecija. . The Philippine Entomologist. 2017 Oct; 31(2):151–52.

[2] Roque, A. Armyworms cut Nueva Ecija onion output; imports loom. Philippine Daily Inquirer [Internet] 2018 Apr 6 [cited 2020 Apr 7]; Available from: https://newsinfo.inquirer.net/

[3] Montecalvo M, Navasero M. Susceptibility of onion armyworm, Spodoptera exigua (Hübner) (Lepidoptera: Noctuidae), larvae to Spodoptera exigua multiple nucleopolyhedrovirus (SeMNPV). Journal of the International Society for Southeast Asian Agricultural Sciences. 25(2): 23-30.

[4] Rosenzweig C, Iglesius A, Yang XB, Epstein PR, Chivian E. Climate Change and Extreme Weather Events - Implications For Food Production, Plant Diseases, And Pests . Global Change and Human Health. 2001 Dec; 2(2):90–104.

[5] Witten IH, Frank E, Hall MA. Data Mining: Practical Machine Learning Tools and Techniques. 3rd ed. Burlington, MA: Morgan Kaufmann; 2011.

[6] Arulkumar G, Manisegaran S, Nalini R, Mathialagan M. Seasonable abundance of beet armyworm Spodoptera exigua (Hübner) infesting Onion with weather factors in Madurai district of Tamil Nadu. Journal of Entomology and Zoology Studies. 2017; 5(6):1157–62.

[7] Calderon RA, Damian MAE, Dumlao MF, Ambat SC, Dela Cruz GB. Determining Bug Epidemic Patterns Using Decision Tree Algorithm. International Journal of Engineering Research and General Science [Internet]. 2016; 4(1):118–27.

[8] Balaban İ, Acun F, Arpalı OY, Murat F, Babaroğlu NE, Akci E, et al. Development of a Forecasting and Warning System on the Ecological Life-Cycle of Sunn Pest. 2019 [cited 2020 May 9]; Available from: https://arxiv.org/abs/1905.01640

[9] Kotsiantis S, Kanellopoulos D, Pintelas PE. Data Preprocessing for Supervised Learning. International Journal of Computer Science. 2006; 1(1):111–17.

[10] Geron A. Hands-On Machine Learning with Scikit-Learn & TensorFlow. Second Edition. Sebastopol, Canada: O'Reilly Media Inc.; 2019.

[11] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research. 2002 Jun 1; 16 :321–57.

[12] Corrales DC, Corrales JC, Figueroa-Casas A. Toward detecting crop diseases and pest by supervised learning. Ingenieria y Universidad. 2015; 19(1):207–28.

[13] Breiman L. Machine Learning. 2001; 45(1):5–32. doi:10.1023/a:1010933404324

[14] Hastie T, Friedman J, Tisbshirani R. The elements of Statistical Learning: Data Mining, Inference, and prediction. New York: Springer; 2017.

[15] Kuhn M, Johnson K. Applied predictive modeling. New York: Springer; 2013.

[16] Elmurngi E, Gherbi A. An empirical study on detecting fake reviews using Machine Learning Techniques. 2017 Seventh International Conference on Innovative Computing Technology (INTECH). 2017 Aug; 107–14. doi:10.1109/intech.2017.8102442