Proceedings of the Second

University Journal of Research and Innovation

August, 2020

Organized by
University of Computer Studies (Pakokku)

Proceedings of

The Second University Journal of Research and Innovation 2020

August, 2020

Organized by

University of Computer Studies (Pakokku)

Department of Higher Education,

Ministry of Education, Myanmar

University Journal of Research and Innovation

Volume 2, Issue 1 2020

University Journal of Research and Innovation 2020

Volume 2, Issue 1, 2020

This journal and individual paper published at www.ucspkku.edu.mm.

Copies:110

All research papers in this journal have undergone rigorous peer-reviewed which is published annually. Full papers submitted for publication are refereed by the Associate Editorial Board through an anonymous referee process.

The authors of the paper bear the responsibility for their content.

Papers presented at the Second University Journal of Research and Innovation (UJRI), University of Computer Studies (Pakokku), August 2020.

UJRI 2020 Editorial Board

UJRI 2020 Editorial Board

Editor in Chief

Proceedings of

The Second University Journal of

Information and Computing Science 2020

Augest, 2020

Contents

Artificial Intelligence & Machine Learning

Big Data Analysis

Data Mining & Machine Learning

Database Management System & Information Retrieval

Digital Signal Processing

Embedded System

Image Processing

Network & Security

Software Engineering and Web Engineering

Machine Learning Based Web Documents Classification

Myat Kyawt Kyawt Swe
Faculty of Computer Science, University of Computer Studies (Hpa-an)
myatkyawt1229@gmail.com

July Lwin
Faculty of Computer Science, University of Computer Studies (Hpa-an)
julylwin7788@gmail.com

Abstract

Nowadays, the research area of Web documents classification based on Machine Learning Approach becomes important filed of research. There is important problem that need to categorize these documents from the Web based on predefined categories. Because the huge amount of Web documents becomes more difficult to effectively discover the target information for the online users. In recent year, Support Vector Machine (SVM) and Convolutional Neural Network (CNN) have been used in many classification research areas and the better performance result can be achieved by using these classifiers. For this system, two popular classifiers from machine learning approaches are used to get the better performance of the system based on the different amount of training dataset. The performance result of the system will be shown with the various analysis on the different trained dataset. The accuracy results of CNN classifier is better than SVM classifier.

Keywords: Convolutional Neural Network, Support Vector Machine, Machine Learning, Web Document Classification, Document Retrieval;

1. Introduction

In the recent year, according to the improvement of user on the Internet, the number of documents on Web are growing with and exponential rate and huge amount of documents can also make more and more difficult to effectively observe the important information for online user on Internet. Therefore, the automatic classification for Web documents improve as very important place in many research areas, that are operated on the huge amount of dataset to get better results.

There is a huge amount of information Web documents that is improving at an exponential rate every day in the form of news articles. These large amount of Web documents can be found under the Website and these documents are linked to each other using various ways. Machine Learning based on Web documents classification becomes an important problem to achieve more and more important day by day. Support Vector Machine and Convolutional Neural Network are supervised learning classifiers from Machine Learning approaches and then these classifiers consists main two part (training part and testing part).

An effective classification model can be developed with a certain amount of labeled samples to solve the classification problem. This proposed model is built using Machine Learning approaches for the users to divide the categories of Web documents which users want to about their interesting information. As the dataset of the system, the Web documents related to the five categories (e.g. Health, sport, education, business and weather) are collected to build the dataset of proposed system. After passing the preprocessing step of this system, classification process is performed and the trained dataset and classifiers are used to predict the category of output of the input test documents. The Machine Learning classifiers, SVM and CNN have been implemented and tested on the various training dataset in the various training dataset in the system's classification step.

2. Related Works

In the previous years, there have been rapid progresses and extensive investigations in automatically hierarchical classification. M.Y. Kan and H. O. N. Thi have developed four structures used in text classification: virtual category tree, category tree, virtual directed acyclic category graph and directed acyclic category graph [1]. In their work, the models depend on the hierarchical structure of the dataset and the assignment of documents to nodes in the structure. Nevertheless, most of researchers adopt a top-down model for classification.

In 2012, F. Sebastiani found a boost of the accuracy for hierarchical models over flat models using a sequential Boolean decision rule and multiplicative decision rules [2]. They used in many Web directory service since the documents could be assigned to both internal and leaf categories. These two models used category-similarity measures and distance-based measures to describe the degree of falsely classification in analyzing the classification performance.

There are many studies developed for this purpose in the literature. Automatic Web documents classification is a problem that can be solved with Supervised Learning approaches. There are many studies developed in the literature using supervised algorithms for web documents classification. Some of the traditional machine learning algorithms commonly used in the literature are as follows; Decision Trees, Artificial Neural Network, Support Vector Machines, K Nearest Neighbor, Bayesian Algorithm. In addition to these algorithms, studies on classification of web pages based on Genetic Algorithm, Ant Colony algorithm are also used in literature [3].

Although the Web documents classification problem has some structural differences from the conventional Web documents classification problems, the algorithms and approaches used in the classification process are similar to each other. Studies on Web documents classification in the literature have been examined, before the development of a system for the classification of Web documents within this study. As with many machine learning approaches have begun to be used widely in the field of Web documents classification in recent years. Web documents classification approaches developed using machine learning approaches are very important in the literature.

3. Background Theory

In this system, the huge amount of Web documents are collected form the various Websites and then these documents are categorized into the predefined categories using two classifiers (SVM and CNN) from the Machine Learning approaches. The SVM and CNN are very useful classifiers for the many classification research areas.

3.1. Support Vector Machine

One of the Supervised Machine Learning Model is Support Vector Machine (SVM) that uses classification algorithm as two groups' classification problems. After creating the SVM model with the labeled training data for each predefined category. In SVM model, training dataset is required that are associated labels with it [4].

For some real world application, tens or even hundreds of features are used in the SVM model. Moreover, thousands of features are used in NLP classifier, when they can have up to one for every word in the training dataset. While using nonlinear kernels may be a good idea for other problems, with this many features will end making nonlinear kernels over fit the dataset. So, it is the best to stick a good previous linear kernel, which really provide the best performance for these problems. SVMs have been developed successfully on various kinds of classification research areas and have consistently implemented better than other non-linear classifier like Neural Network (NN) and Mixtures of Gaussians. The improvement of efficient Web documents classification model forward to the use of SVMs for classification of documents categorization [5].

3.2. Convolutional Neural Network

CNNs are generally applied in computer vision, but they have recently been implemented to various NLP research area and the results were promising. CNN structures are a deep learning approach with an unbiased structure used for NLP problems [6].

CNNs are generalized versions of multi-layer perceptron. Multi-layer perceptron usually provide fully connected networks layers, which is, each neuron of one layer is connected to all neurons of next layer. The "fully-connectedness" of these networks preform them prone to over fitting data. Normal ways of regularization include adding some form of magnitude measurement of weights to the loss function. Convolutional Neural Network classification models take a different approach towards regularization: they take advantage of the hierarchical pattern in data and assemble more complex patterns using smaller and simpler patterns.

4. Methodology of Proposed System

The main purpose of the proposed system is to categorize the output class to the documents articles based on the predefined label of contents, and figure 1 describes the flow of the proposed system design. In the initial step, to collect the dataset of the system, Data Retrieval module where in self-development web scraping algorithm is performed to extract the actual text from the Web documents. Then the effective features are extracted from this document to get the useful dataset.

To create the effective training dataset of the system, SVM and CNN classifiers are used in this system and the better performance results are achieved. The performance results are evaluated using Precision, Recall and F-measure based on the two classifiers with various training dataset sizes. The accuracy result of the system can be changed based on the amount of training dataset because the supervised learning approaches are used in this system.

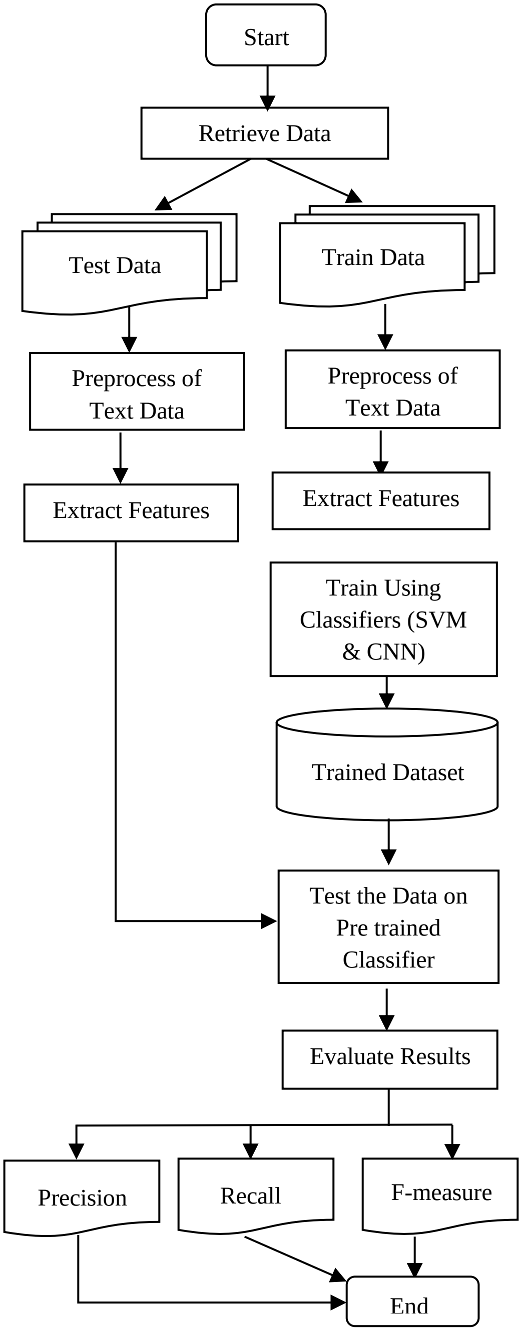

Figure 1. The Proposed System Design

5. Results and Analysis

The model that has been created is necessary to evaluate in order to be able to see its performance in performing the tasks ordered. There are several ways of evaluation used in the model.

5.1. Data Collection

In the step of the data collection, the related documents are collected from the International News and Asia News Website (https://www.channelnewsasia.com/). The collected dataset must be divided into two main parts (training dataset and testing dataset) to create the supervised learning model. For this system, 75% of total dataset is used to create the training dataset and 25% of total dataset is used for testing dataset. For the training step, data preprocessing methods are used to train the dataset as an input of the classifier. And then the same preprocessing methods of data are also used for testing data and this acts as an input to trained classifier which predicts the output class of the test documents.

5.2. Preprocessing

In this phase, tokenizing the articles is the first step of the system that is used to convert the characters sequence into the string sequence with predefined meaning. After tokenizing, stemming the words is performed on the every word of news articles to reduce related inflectional form. Then the stop words are removed as the important words and more significance words as main process of classification. In this step, specific stop words list are built to remove the stop words.

5.3. Performance Analysis

In this paper, accuracy, precision, recall and F-measure are used to evaluate the performance of the classifier. Prior to the evaluation phase of the system, the results of the classification process are categorized into four types, namely true positive (TP), true negative (TN), false positive (FP), false negative (FN).

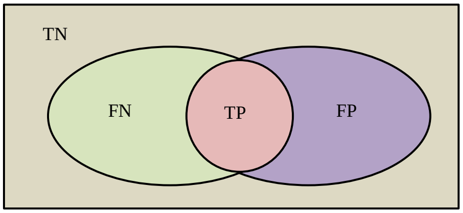

Figure 2. The set of evaluation parameters

TP or true positive is the amount of positive data that is classified into positive class. TN or true negative is the amount of negative data that is classified into negative class. FP or false positive is the amount of positive data classified to the negative class. FN or false negative is the amount of negative data classified to the positive class. In particular, the calculation of these measurement parameters used in data extraction is a given in figure 2.
The accuracy of classification is calculated by the following formula:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \qquad (1)$$

Precision is a measure of the accuracy between the information requested by the user and the answers given by the system. The difference between accuracy and precision is, the accuracy shows the proximity of the measured result to the true value, precision indicating how close the difference in value when the repetition of the measurement takes place. It is computed as in equation (2).

$$Precision = \frac{TP}{TP+FP} \qquad (2)$$

In the same fashion, Recall is the measure to know fraction of correctly classified instances among all the instances that are supposed to be in correctly labelled instances. It is measured as in equation (3).

$$Recall = \frac{TP}{TP+FN} \qquad (3)$$

There is another measure which is based on the precision and recall. It is harmonic mean of these two measures. It is called F-Measure which is computed as in equation (4).

$$FMeasure = 2 \times ((Precision \times Recall)/(Precision + Recall)) \quad (4)$$

The performance analysis of the system is evaluated on the various training dataset size (500, 1000 and 1500 training documents). The accuracy results can be changed based on the size of training dataset of the system. Table 1 shows the changes of accuracy result on the various training dataset size. The accuracy result highly depends on the amount of training dataset because the system used supervised learning algorithm.

Table 1. Accuracy Result on Various Dataset

Algorithm	**Trained 500 dataset**	**Trained 1000 dataset**	**Trained 1500 dataset**
SVM	75.2	82.5	90.4
CNN	78.56	86.2	94.5

The graph of the analysis of the accuracy results are shown in the figure 3.

The blue bars represent the accuracy results of the SVM classifier and the red bar represent the accuracy results of CNN classifier. According to the analysis, the accuracy of these classifiers are improved based on the increasing the amount of trained dataset and then the better accuracy results are achieved using CNN classifiers than the SVM classifier.

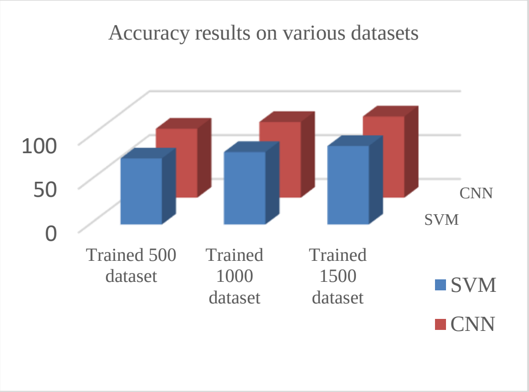

Figure 3. Accuracy Analysis

Moreover, the experimental results of the system are measured based on the Precision, Recall and F-measure. These performance results are calculated on the trained dataset 1500 documents. All performance results of the CNN classifiers can be provided the better results than the SVM classifiers. Table 2 shows the performance results of the system. The values of performance measurement on the table 2 are calculated based on the total 1500 trained dataset.

Table 2. Performance results on 1500 trained dataset

Classifiers	**Precision**	**Recall**	**F-measure**
SVM	86.5	88.7	89.5
CNN	88.2	90.1	92.2

According to the performance analysis in table 2, the value of performance measure of CNN classifier is better than SVM. The step by step layer of CNN, the 15 layers architecture of convolutional layers and max pooling layers can provide the more exact classification result of the system.

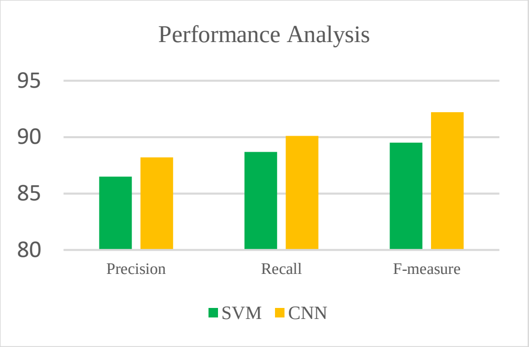

Figure 4. Performance Analysis

The performance bar chart is shown in figure 4 with the two color bar. The green shows the performance results of SVM classifier and the yellow bars represents the performance results of the CNN classifiers.

6. Conclusion

In this paper, the Web documents classification model is created based on the supervised learning methods of Machine Learning approaches. The effective condition of CNN is used in the creating the training dataset model and prediction of the output result of the system. The accuracy results of the system are analyzed based on the various training dataset seize. The lager amount of trained dataset is provided the better result of the classification based on the supervised learning approaches. The performance evaluation results of the two classifiers are analyzed. According to the evaluation results the performance results of the Convolutional Neural Network classifier provide the better result than the Support Vector Machine classifier.

References

[1] M.Y. Kan and H. O. N. Thi. "*Fast webpage Classification using URL features*". 14th ACM, pages 325–326, New York, NY, 2005. ACM Press.

[2] F. Sebastiani. "Machine learning in automated text categorization". ACM Computing Surveys, 34(1):1– 47, March 2002.

[3] N. Holden and Alex A. Freitas, "Web Page Classification with an Ant Colony Algorithm", LNCS, Springer, 2004, Vol.3242, pp:1092-1102.

[4] D. Lewis "Naive (bayes) at forty: The independence assumption in information retrieval," Machine Learning: ECML-98, pp. 415, 1998

[5] J,.K. M. an, "Data Mining: Concepts and Techniques", 2nd ed. 2006.

[6] R. Collobet and et al., "Natural language processing (almost) from scratch". JMLR 12:2493–2537.

COVID-19 Threat Prediction with Machine Learning

Myat Thet Nyo
University of Computer Studies (Meiktila), Myanmar
myatthetnyo05@gmail.com

Aye Aye Naing
University of Computer Studies (Banmaw), Myanmar
aanaing85@gmail.com

Yi Yi Win
University of Computer Studies (Meiktila), Myanmar
yiyipku941@gmail.com

Abstract

The outbreak of SARS-CoV-2 also known as COVID-19 coronavirus has become a big threat to living society. The whole world faces several bottle necks caused by COVID-19 and all are trying to fight against the spread of this pandemic. Machine learning is the powerful tool to analyze the track of disease, predict the growth of the epidemic, and also provide an effective help to design policies to down its spread. This study applies linear regression and polynomial regression to predict the threat of COVID-19 in Myanmar. We conclude that polynomial regression gains less error rate and more prediction accuracy.

Keywords: Machine Learning, Regression, COVID-19, SARS-CoV-2, Prediction

1. Introduction

COVID-19 is a respiratory infection with common signs that include respiratory symptoms, fever, cough, shortness of breath, and breathing difficulties. In more severe cases, infection can cause pneumonia, severe acute respiratory syndrome, kidney failure, and death. China first reported the Coronavirus disease on 31st December 2019 in Wuhan. Shortly it started spreading local transmission, communication transmission and rapidly across the world. On January 30, 2020, the World Health Organization (WHO) declared this outbreak a Public Health Emergency of International Concern (PHEIC) as it had spread to 18 countries. On Feb 11, 2020, WHO named this “COVID-19”. On March 11, as the number of COVID-19 cases has increased thirteen times apart from China with more than 118,000 cases in 114 countries and over 4,000 deaths, WHO declared this a pandemic.

Based on the globally shared live data by the WHO coronavirus Disease (COVID-19) Dashboard, there are 8,385,440 confirmed cases, out of which 450,686 cases passed away [1]. Based on the locally shared live data by the Ministry of Health and Sports, Myanmar, there are 286 confirmed cases, out of which 197 cases are recovered, and out of which 6 cases lost their lives.

As the outbreak becomes pandemic, there is a need to develop, analyze the data on the spreading diseases and to provide an effective way to fight the disease. Machine learning methods: linear regression and polynomial regression is applied in this study to predict the spreading of COVID-19.

2. Related Works

A simple machine learning approach was developed in [2] to forecast the range expansion of 2009 H1N1 flu pandemic. The future distribution of the target species was predicted by composing the spreading patterns of the previous range expansions of various alien species similar to the target species. A matrix representing the early-late relationship of infestation between regions was used to compare and compose various range expansion patterns. The study area was divided into many small regions, and the dates of infestation in the regions were predicted. Their method successfully predicted the infestation dates of various alien species and the novel H1N1 influenza pandemic in 2009.

Logistic regression and gradient boosted trees methods are used in [3] to predict the severe complication of COVID-19. Machine learning models were created that use a patient's historical medical data to check pneumonia, influenza, acute bronchitis, and other specified upper respiratory infections. Logistic regression is aimed for identifying the individuals who are at risk. They identify risk features as: older adults, individuals with heart disease, individuals with diabetes, and individuals with lung disease. They created two variations of the model of gradient boosted trees. The first is a model that leverages information similar to their logistic regression model and the next one is full diagnosis histories to be leveraged within XGBoost model.

Robust Weibull model based on iterative weighting was proposed in [4]. Using iterative weighting for fitting Generalized Inverse Weibull distribution, a better fit was obtained to develop a prediction framework. For the base line, the gaussian distribution was deployed to estimate the number of cases with time. Five sets of global data on daily new COVID-19 cases were used to fit parameters of five different distributions. They concluded that their proposed function fits the best to the COVID-19 dataset.

3. COVID-19 Myanmar

According to Myanmar's Ministry of Health and Sports (MOHS), 56,726 people had been tested for the coronavirus, with 56,463 testing negatives. From 31st January to 18 June, Myanmar quarantined 85 people at hospitals and 41,304 people at other quarantine centers. Of the 263 positives cases, 187 people are recovered and 6 people are dead [5].

According to Myanmar's Ministry of Health and Sports (MOHS), 103 people of total confirmed case are imported and 160 people are local transmission cases. Confirmed cases were detected as a peak in the 4th week after first confirmed cases due to the religious event cluster. Confirmed cases were decreased but still raised than the first 3 weeks due to the religious event cluster and workplace cluster. There were an increased in numbers of cases in the 7th week after the first two cases due to the household cluster linked with religious cluster. And confirmed cases were detected by n nearly two-fold rise in the 9th and 10th weeks after the first two cases because of returnees coming back by relief flights.

3.1. Dataset

The day to day prevalence data of COVID-19 from January 2020 to 18 June 2020, were retrieved from the official Coronavirus Disease 2019 (COVID-19) Surveillance Dashboard (Myanmar). The dataset is composed of daily case reports and saturation reports Ministry of Health and Sports with four attributes i.e. state, confirmed, death, and recovered cases, where the update frequency of the dataset is once in a day.

The bar charts of total confirmed cases and total dead cases of Myanmar are shown in figure 1 and 2.

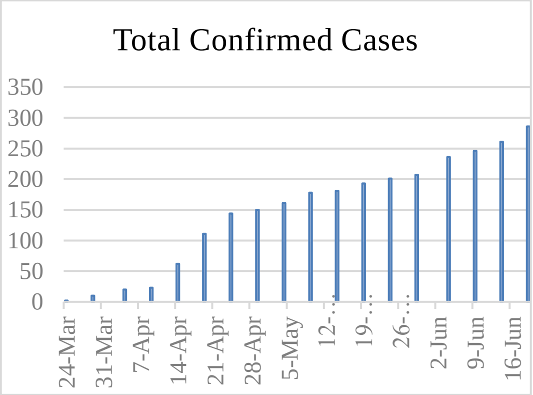

Figure 1. Total Confirmed Cases of Myanmar

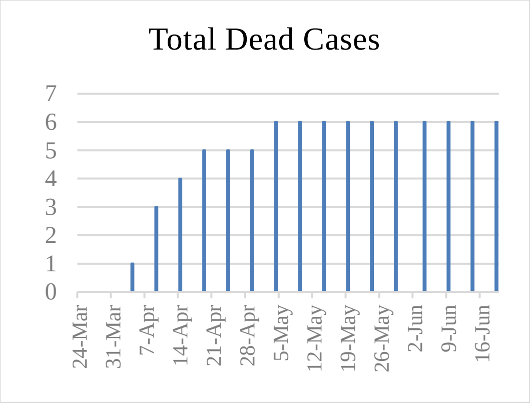

Figure 2. Total Dead Cases of Myanmar

4. Method

We apply two machine learning methods: linear regression and polynomial regression for comparison of performance on prediction.

4.1. Linear Regression

Linear regression is a basic and commonly used type of predictive analysis which usually works on continuous data [6]. Linear regression consists of finding the best-fitting straight line through the points. The best-fitting line is called a regression line. This type method observes the data and comes to the conclusion that the data is linear after plotting the scatter plot. As soon as a pattern is saw in the data, it plans to make a regression line on the graph so that can use the line to predict the next data. Using the training data, a regression line is obtained which will give the minimum error so that it is able to predict the required output.

$$Yi = \beta_0 + \beta_i Xi + \varepsilon_i \qquad (1)$$

Here, the β1 it's the parameters (also called weights) βo is the y-intercept and ε_i is the random error term whose role is to add bias. The above equation is the linear equation that needs to be obtained with the minimum error. It can be simplified as

Y (predicted) = (β1*x + βo) + Error value (2)

Where '**β1**' is the *slope* and '**βo**' is the *y-intercept* similar to the equation of a line. The values 'β1' and 'βo' must be chosen so that they minimize the error.

$$Error = \sum(Actual \; - Predicted\,)^2 \qquad (3)$$

We take total confirmed cases, recovered cases and dead cases as attributes into account and forecast the confirmed cases and dead cases. The prediction of confirmed cases with linear regression vs actual confirmed cases is shown in figure 3. And the prediction of dead cases with linear regression vs actual dead cases is shown in figure 4.

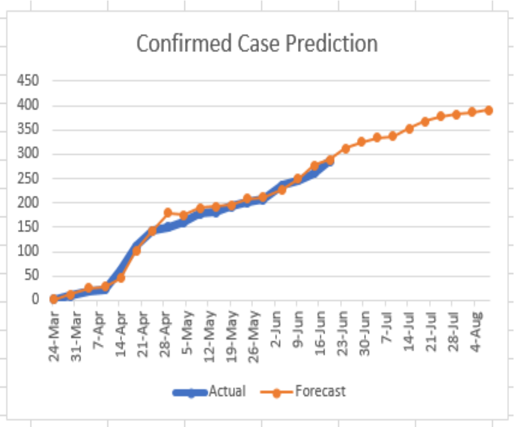

Figure 3. Confirmed Case Prediction with Linear Regression

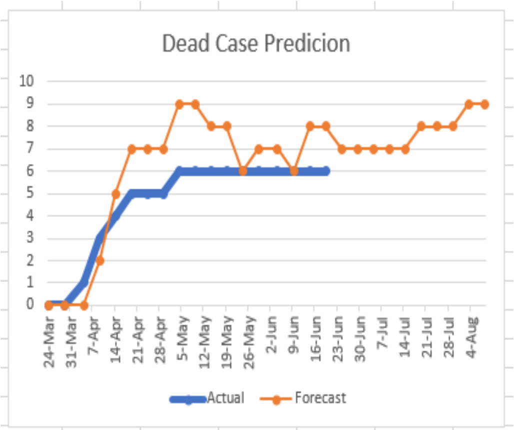

Figure 4. Dead Case Prediction with Linear Regression

4.2. Polynomial Regression

It is very difficult to fit a linear regression line with a low value of error. Hence, we can try to use the polynomial regression to fit a polynomial line so that we can achieve a minimum error or minimum cost function. This is one of the regression techniques which is used by the professionals to predict the outcome. It is defined as the relationship between the independent and dependent variables when the dependent variable is related to the independent variable having an nth degree. The equation of the polynomial regression would be:

$$Yi = \beta_0 + \beta_i Xi + \beta_i Xi^2 + \varepsilon_i \qquad (4)$$

Here, the β1 it's the parameters (also called weights) βo is the y-intercept and ε_i is the random error term whose role is to add bias as in equation 3.

` Polynomial provides the best approximation of the relationship between the dependent and independent variable. A broad range of function can be fit under it and it basically fits a wide range of curvature [6].

We take total confirmed cases, recovered cases and dead cases as attributes into account and forecast the confirmed cases and dead cases. The prediction of confirmed cases with polynomial regression vs actual confirmed cases is shown in figure 5. And the prediction of dead cases with polynomial regression vs actual dead cases is shown in figure 6.

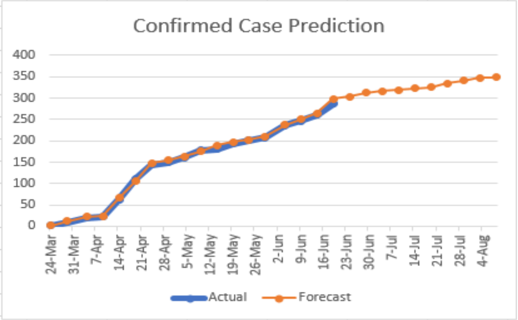

Figure 5. Confirmed Case Prediction with Polynomial Regression

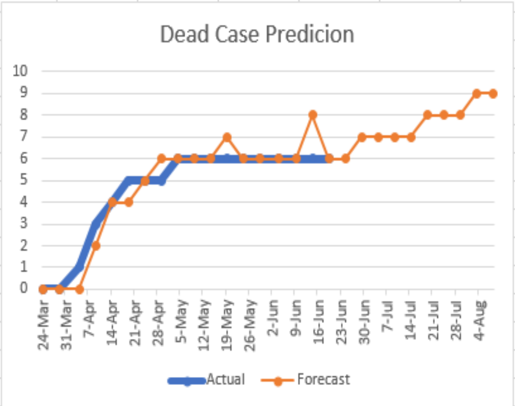

Figure 6. Dead Case Prediction with Polynomial Regression

4.3. Performance Comparison

The mean squared error (MSE) is the most widely used objective function and root mean square error (RMSE) as a metric function for evaluating the regression models. The MSE loss can be computed by using equation (5).

$$L(y, y^\wedge) = \frac{1}{N} \sum_{i=0}^{N} (y - y^\wedge) \qquad (5)$$

Where y indicates the original value, ^y indicates the predicted value, and N is the number of samples predicted.

According to the results of prediction with linear regression and polynomial regression, the RMSE values are shown in table 1.

Table 1. Comparison of Linear and Polynomial Regression

Method	Confirmed	Dead
Linear	3.2	1.3
Polynomial	2.8	1.04

The comparison chart of actual confirmed cases and predicted cases with polynomial and linear regression is shown in figure 7. And the comparison chart of actual dead cases and predicted cases with polynomial and linear regression is shown in figure 8.

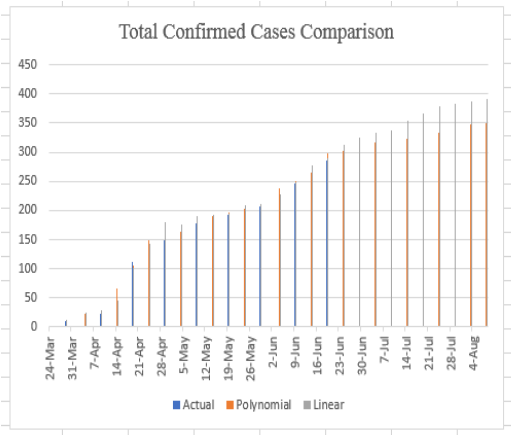

Figure 7. Confirmed Case Prediction Comparison with Polynomial and Linear Regression

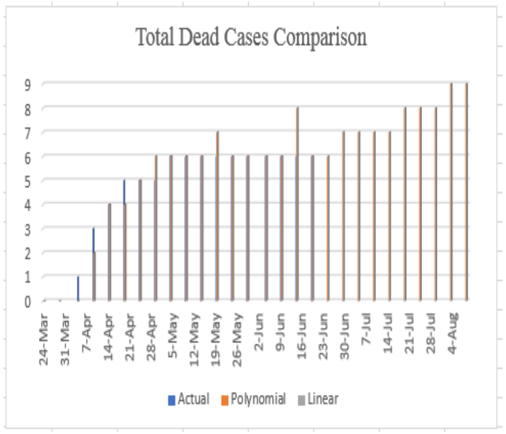

Figure 8. Dead Case Prediction Comparison with Polynomial and Linear Regression

5. Conclusion

COVID-19 is still an attacking threat to the living society and the researcher of several fields are trying to fight back in their own ways. Prediction is a better way to manage the pandemical situations and to be able to make plans to take actions. We applied two machine learning methods, linear regression and polynomial regression and take a performance comparison using RMSE units. We conclude that polynomial regression gains less RMSE units in prediction of COVID-19 confirmed cases and dead cases.

References

[1] World Health Organization COVID-19 disease Dashboard, https://covid19.who.int/

[2] F.Koike, N.Morimoto, "Supervised forecasting of the range expansion of novel non-indigenous organisms: Alien pest organisms and the 2009 H!N! flu pandemic", Global Ecology and Biogeography Journal, John Wiley & Sons Ltd 2018.

[3] D.Decapprio, J.Gartner, C.J.McCall, "Building a COVID-19 Vulnerability Index", March 24, 2020.

[4] T.Shreshth, T.Shikhar, T.Rakesh, S.S.Gill, "Predicting the growth and trend of COVID-19 pandemic using machine learning and cloud computing", Internet of Thing 11, 2020.

[5] The Republic of the Union of Myanmar, Ministry of Health and Sports, "COVID-19 Surveillance Dashboard(Myanmar)", www.mohs.gov.mm

[6] A.Pant, "Introdcution to Linear Regression and Polynomial Regression", Jan 13, 2019, www.towardsdatascience.com

University of Computer Studies (Pakokku)
Department of Higher Education
Ministry of Education
Myanmar