Sentiment Analysis using Logistic Regression

George B. Aliman, Tanya Faye S. Nivera, Jensine Charmille A. Olazo, Daisy Jane P. Ramos, Chris Danielle B. Sanchez, Timothy M. Amado, Nilo M. Arago, Romeo L. Jorda Jr., Glenn C. Virrey, Ira C. Valenzuela

Abstract

This paper proposed a study that will assess different machine learning techniques in classifying tweets. There are four machine learning techniques that will be subjected to testing using same set of data namely: Naive Bayes, Linear Support Vector Classifier, Stochastic Gradient Descent Classifier and Logistic Regression. It is always a challenge to identify which machine learning model will give the most efficient performance in sentiment analysis. The main objective of this paper is to find the best machine learning technique for the sentiment analysis in English, Filipino and Taglish languages. The said models will be integrated to Twitter's API for the collection of twitter data which will be subjected to data preprocessing to make the tweets analyzable and then feature extraction was done using Natural Language Processing. The performance scores of each machine learning algorithm has been computed. The four algorithms: Support Vector Classifier, Stochastic Gradient Descent, Naive Bayes and Logistic Regression were used for machine learning with an accuracy of 69%, 71%, 77%, and 81% respectively. The Logistic Regression Model has the highest accuracy and best fitted algorithm for prediction of potential mental health crisis tweets.

***Keywords*— Sentiment Analysis, Machine Learning Algorithms, Twitter, Tweepy, Mental Health Crisis**

I. Introduction

Sentiment analysis inspect people's opinions, sentiments, evaluations, attitudes, and emotions through written language [1]. It has been very useful in almost every business and social domain, because the comments and reviews form the people determines the emotion of a customer towards their products and services [2]. Many methods such as Natural Language Processing (NLP) is used to perform sentiment analysis. NLP lets computers learn, read, and regulate the language of humans.

Machine Learning (ML) is a collection of techniques and algorithms used to design systems that acquire from multiple data [3]. It can perform predictive analytics far faster than any human can do. Machine Learning can help human work more efficient as possible. There are dozens of machine learning algorithms that are supervised and unsupervised, and each takes a different approach to learning [4] [5] [6] [7]. The most commonly used techniques in machine learning are Naive Bayes, Linear Support Vector Classifier, Stochastic Gradient Descent Classifier, and Logistic Regression. Naive Bayes Algorithm is a strong method of machine learning used in classification tasks and predictive modeling [8]. It can very readily be written into the code providing model predictions in very little time. Naïve Bayes is simple and quick to predict sample information set class as well as performing well in multi-class prediction. Support Vector Machine (SVM) is one amongst the essential algorithms of pattern recognition [9] that is based on a non-probabilistic linear classification. Primarily SVM takes the pattern that are linearly severable and moves them into a hyperplane space. The patterns that do not seem to be linearly separable are reworked into a new space and mapped with original information by utilizing kernel function. Stochastic Gradient Descent (SGD) Classifier is an algorithm that is efficient to use in minimizing the cost function by finding the values of coefficients of a function [10]. In using a very large dataset, it is more efficient to use the SGD classifier than the Gradient Descent (GD) optimization. Logistic Regression is an algorithm used in classification tasks and predictive analysis. It uses linear regression equation to make discrete binary outputs but unlike linear regression, its cost function is Sigmoid function [11]. This function is an S-shaped curve and can also be called the logistic function [12]. The hypothesis of this algorithm tends to limit the logistic function between 0 and 1.

The collection of data has occurred at its peak, and the amount of information in the world has been estimated to double every two years [13]. With this amount data,

Technological University of the Philippines-Manila
Manila, Philippines

data mining is used in identifying the patterns of huge and complex data sets. However, a data source like Twitter which produces an increasing amount of data every day, a problem in classification also emerges in data mining. Since no best technique is available or one size fits all, this study will assess different algorithms to find the right algorithm based on testing the accuracy using the same datasets.

The attention of the study is focused on Twitter as the source for data mining because of its popularity and the huge amount of data it generates. The objective of this study is to collect and analyze twitter data to classify its sentiment using Natural Language Processing and Machine Learning Algorithm such as Naive Bayes, Linear SVC, SGD Classifier and Logistic Regression for classifying tweets. These Machine Learning Models will be compared according to its performance in classifying sentiments. Most efficient ML model will be used in the system for sentiment analysis in the three languages namely: English, Filipino and Taglish. The accuracy, f1 score, precision, and recall will be computed from the output of the different algorithms used in sentiments analysis [14] [15] [16] [17].

Finding the best classifier is very significant since it will be utilized in classifying the emotion of the Twitter users. Evaluating the sentiment of a person should be done in most careful and sensitive way. The machine learning model to be used must have a high accuracy score.

II. Related Studies

I. Rish presented an empirical study about Naïve Bayes [18], which will help people to understand the data characteristics that affect Naive Bayes performance. Bayesian classifiers assign the most likely class to a given example described by, $P(X|C) = \prod_{i=1}^{n} P(X_i\,|C)\ where\ X = (X_1, ..., X_n)$ is a feature and C is a class. Simulations from Monte Carlo was used which enabled a systematic study of classification accuracy for several classes of randomly generated problems. The effect of distribution entropy on classification error was evaluated, showing that distributions with low entropy features yield good output with Naïve Bayes. It is shown that naive Bayes works well for some near-functional feature dependencies, thus achieving its best output in two opposite cases: completely independent as expected and surprisingly a functionally dependent feature. As a result, Naive Bayes accuracy is not specifically associated with degree of feature dependencies calculated as class conditional reciprocal knowledge between characteristics. Instead, a better measure of naïve Bayes accuracy is the amount of class knowledge that is lost because of the presumption of independence.

In order to reduce the damage caused by depression, a portable and accurate depression detection and diagnosis method is conducted by Shen et al. in 2017 [19]. They presented a method for pervasive electroencephalogram (EEG) based detection and used a three-electrode pervasive EEG collection device for diagnosis of depression. In this study, Support Vector Machine was utilized for classification problems by letting $D= \{(x_i, y_i) \,|\, x_i \in R_n, y_i \in \{-1, +1\}, i = 1, 2, ..., m\}$ where y_i is the label of $x_{i.}$ The EEG data collected through the device from 170 subjects have been analyzed using Support Vector Machine and the accuracy reaches 83.07%. The results show that the method used for detection and diagnosis of depression is suitable and effective. It also indicates that SVM is an effective model in analyzation and classification.

The paper proposed by Mittal et al. [20], combined an unsupervised Artificial Neural Network (ANN) method named Self -Organizing Maps (SOM) with a supervised classifier called Stochastic Gradient Descent (SGD) in diagnosing breast cancer. Initially the SGD approach is used in isolation for the classification function, and after hybridization with the unsupervised ANN technique on the Wisconsin Breast Cancer Database (WBCD) to perform the classification. The output classification by SGD is given as $c(x) = v^T + j$ where v belongs to R^m and j being the intercept which belongs to R (system for regularization). In addition, the findings of the classification experiment using SOM hybridization with SGD are much superior to SGD. By integrating SGD with self-organizing maps, the hybrid model built up gave a high precision value over training set and test set. The consistence of the accuracy of the algorithm was tested using a much bigger one dataset.

In a study by Peduzzi et al. [21], a Monte Carlo was performed to assess the effect of the number of events per variable (EPV) analyzed in the logistic regression analysis Deaths and survivors were tested separately, based on the expected risk of dying (Pi) or surviving (Q = I – P) the logistic model, where Pi = 1/{1 + + Xl þ)]}; is the intercept term; Xi = (Xl l, ,Xi 7) is the set of covariate values for patient i; and = (PI, 4), is the set of the corresponding regression coefficients, calculated from the complete EPV sample = 36.

The simulations were based on data from a 673-patient cardiac trial, in which 252 deaths occurred and seven variables were cogent mortality predictors; the number of incidents per predictive variable was (252/7=) 36 for the full study. For the simulations, at EPV values = 2, 5, 10, 15, 20 and 25, randomly generated 500 samples of the 673 patients, selected with substitution, according to a logistic model derived from the full sample.Overall, when the ratio of the number of events per variable analyzed is high, the validity of the logistic model is problematic.

Machine learning models are proven helpful to increase the classification system for various applications.

III. Methodology

A predictive model was developed using different machine learning algorithms. Figure 1 shows the process flow of the system.

A. Data Collection

Tweepy, an application program interface for Python, was utilized for the collection of Twitter Data [22]. Twitter account has been created to link the Tweepy application using Keys and Tokens [23]. Keywords that has been suggested by the Psychologist was utilized by Tweepy to gather alarming tweets. The search program was set for a specific geolocation and radius for lesser time to be consumed in data mining. Also, the data gathering was set to English, Filipino and Taglish (combination of English and Filipino language) languages.

Table I shows the statistical data gathered with 3,085 positive tweets and 2,310 negative tweets used for training and testing. A positive tweet is a tweet that throws a positive sentiment after analyzing its words and was labeled as "0" in the dataset [24]. On the other hand, a tweet that has been analyzed as negative sentiment was considered as negative tweet. The negative tweets were labeled as "1" in the dataset [24].

Table I.
Statistics of Datasets

Dataset	Positive	Negative	Total
Training	3035	2260	5295
Test	50	50	100

B. Data Preprocessing

The gathered Twitter data has been subjected to data preprocessing to make the text/tweets analyzable [25]. The preprocessing includes the removal of Uniform Resource Locator (URL) from tweets, slang words and avoiding misspelled words. The mention of other accounts with the "@" sign were also removed.

C. Feature Extraction

Some of the machine learning algorithms were created using statistical methods. Natural Language Processing was utilized to extract features from the Twitter data that will be used in Machine Learning [26]. Extraction of the feature was achieved by removing the punctuation mark, word tokenizing, removing the words to avoid, and marking sections, and building frequency distribution. All features extracted were randomized to remove all biases. Feature extraction has then assigned vector values to the preprocessed data and these has been used as a training dataset and testing dataset. A set of words has been given by the Philippine Psychology Association to determine if the posted sentiments in the twitter are posing for possible mental health crisis. Previous posts has been crawled also to further check if the current post shows a mental health crisis.

D. Machine Learning Model Evaluation

All the processed data were fed in different Machine Learning Algorithm such as Naive Bayes, Linear Support

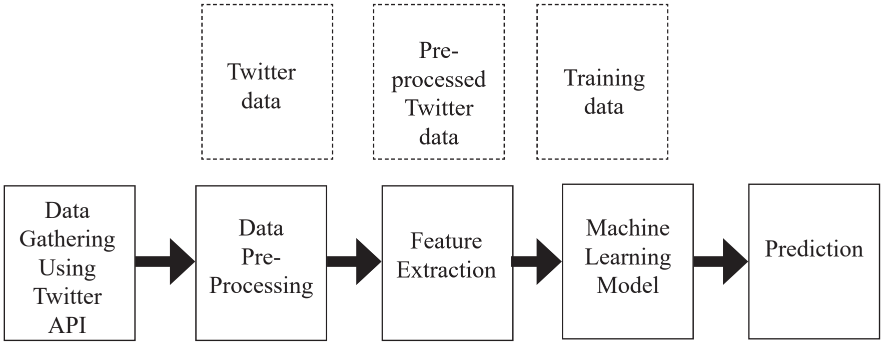

Fig. 1. Process Flow Of The System

Vector Classifier, Stochastic Gradient Descent Classifier and Logistic Regression [27] [28] [29] [30]. These Machine Learning Models have been evaluated according to its performance in classifying sentiments.

Different metrics such as accuracy, precision, recall and f1 score was utilized [14] [15] [16] [17] . These four (4) metrics would tell the effectiveness of the predictions by the system and how a dataset was being trained correctly . Also, it was useful in selecting the best predictive model in classifying tweets.

Equations 1, 2, 3, and 4 was used as statistical evaluation criteria of the performance of the machine learning techniques. Accuracy in Equation 1 measured the ratio of the correctly classified tweets to the whole dataset . Recall in Equation 2 measured the amount of correctly classified negative tweets. While precision in Equation 3 gave the ratio of the total number of classified negative tweets to the total negative prediction in the data sets. On the other hand, F1-score in Equation 4 consider the precision and the recall; it served as the average harmonic mean of the recall and precision.

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \tag{1}$$

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

$$Precision = \frac{TP}{TP+FP} \tag{3}$$

$$F1\,Score = \frac{2(Recall\ x\ Precision)}{Recall+Precision} \tag{4}$$

Where true positive (TP), correctly identified negative tweets which are negative tweets, FP false positive, incorrectly identified negative tweets which are positive tweets, TN true negative, correctly identified positive tweets which are positive tweets, and FN false negative, incorrectly identified positive tweets which are negative tweets.

IV. Result and Discussion

Table II shows the summary rate of true positive, true negative, false positive, and false negative rate in four (4) different machine algorithms on 100 actual tweets. Based on the results, Support Vector Machine has 69 correct predictions and 31 incorrect predictions while Naive Bayes Classifier has 77 correct predictions and there are 23 incorrectly labeled. Stochastic Gradient Classifier has 71 correct predictions and 29 incorrect predictions Whereas Logistic Regression has 81 correct predictions and only 19 were incorrectly labeled.

Table II.
Prediction Summary Rate of Four Algorithms

Machine Models	TP	FP	TN	FN	Population
Support Vector Classifier	44	17	25	14	100
Naive Bayes Algorithm	46	10	31	13	100
Logistic Regression	48	11	33	8	100
SGD Classifier	40	20	31	9	100

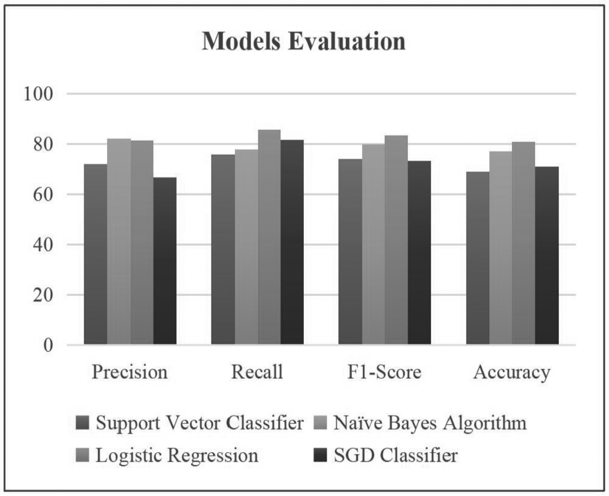

Fig. 2. Performance Measure Of Different Machine Learning Models

Figure 2 shows the graphical representation of performance scores of the four machine learning models. The dataset consists of 5,395 tweets and split into a training set of 5,295 tweets and testing set of 100 tweets to analyze the precision, recall, f1 score and accuracy of different classification models. Based on the results, Logistic Regression performs better than the other models in detection of potential mental health crisis tweets.

Table III.
Performance Measure of Machine Learning Models in Percentage

Machine Learning Model	Precision	Recall	F1-Score	Accuracy
Support Vector Classifier	72.13	75.86	73.95	69
Naive Bayes Algorithm	82.14	77.97	80.00	77
Logistic Regression	81.36	85.71	83.48	81
SGD Classifier	66.67	81.63	73.40	71

Table III shows the precision, recall, f1 score and accuracy of the four machine learning techniques. Naïve Bayes Classifier has the highest precision compared to the other models. Stochastic Gradient Descent has the lowest accuracy with 69% only whereas Support Vector Classifier, Naive Bayes and Logistic Regression obtained an accuracy of 71%, 77%, and 81% respectively. Also, Logistic Regression has the highest recall and f1 score among the four algorithms.

The performance scores indicate that the sentiment analysis model performed well in classifying positive and negative tweet. The best fitted machine learning algorithm was logistic regression since it has the highest accuracy, recall and f1 score compared to the other model and it can be utilized in detecting potential mental health crisis on tweets.

This study focused on collecting and analyzing the sentiments of twitter data. The Twitter user with at least one depressive or negative tweet will be considered as a potential crisis. The system is capable of appeasing the emotions of identified mentally crisis Twitter users by sending motivational and uplifting quotes for the first reply. The bot will also intervene with the user by sending mental health helplines or links for the second reply. The study is to be used only to Twitter users within Metro Manila that are potentially experiencing anxiety and depression based on the tweets. The system can only collect data within seven days prior to the first run of the system. The data gathered was collected at 12 a.m. to 6 a.m. in which mental health crisis attacks are most likely to happen during this time period according to Psychologists. This study is limited to the identification of the reasons for the gathered tweets and to detect the mental health condition of the user.

Comparing the software detection of alarming tweets to the Psychologist assessment is a way to evaluate the accuracy of the system. The system collected fifty (50) alarming tweets from July 24, 2020 to August 1, 2020. Based on Twitter's policy, the proponents take precaution not to reveal the specific username in this paper. For this matter, the username was labeled as "User1" to "User50". Based on the observation, forty-six (46) out of fifty (50) tweets were correctly predicted by the system according to the Psychologist.

V. Conclusion

Twitter is undoubtedly a place where most people express their thoughts and feelings. It is very important to have an appropriate model in prediction of positive or negative tweets. Python programming language was utilized in this study in data mining and developing a predictive model. It used four machine algorithms such as Naive Bayes, Linear SVC, SGD Classifier and Logistic Regression.

In training of machine models, 5,395 data sets of tweets are collected in different Twitter users using Twitter API. The data sets were processed to become a training data and testing data for the machine learning using Natural Language Tool Kit. The machine models' accuracies were evaluated through testing. The four algorithms namely: Support Vector Classifier Stochastic Gradient Descent, Naive Bayes and Logistic Regression were used for machine learning with an accuracy of 69%, 71%, 77%, and 81% respectively. The Logistic Regression Model has the highest accuracy and best algorithm for prediction of potential mental health crisis tweets and is therefore recommended to be used in sentiment analysis.

VI. Acknowledgment

The authors would like to express their deepest and sincere gratitude to the IGAN Team and to the Technological University of the Philippines-Manila for the unwavering support and contributions towards the completion of this research study.

VII. References

[1] H. Saif, Y. He, M. Fernandez and H. Alani, "Semantic Patterns for Sentiment Analysis of Twitter," *The Semantic Web – ISWC 2014, Springer International Publishing,* p. 324–340, 2014.

[2] F.Neri, C.Aliprandi, F.Capeci, M.Cuadros and T.By, ""Sentiment Analysis on Social Media"," *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM),* pp. 919-929, 2012.

[3] P. . Domingos, "A few useful things to know about machine learning," *Communications of The ACM,* vol. 55, no. 10, pp. 78-87, 2012.

[4] A. Aquino, Ma. Veronica Bautista, C. Diaz, I. Valenzuela and E. Dadios, "A Vision-Based Closed Spirulina (A. Platensis) Cultivation System with Growth Monitoring using Artificial Neural Network," *2018 IEEE 10th International Conference on Humanoid, Nanotechnology, Information Technology,Communication and Control, Environment and Management (HNICEM),* pp. 1-5, 2018.

[5] I. Valenzuela, R. Baldovino, A. Bandala and E. Dadios, "Pre-Harvest Factors Optimization Using Genetic Algorithm for Lettuce," *Journal of Telecommunication, Electronic and Computer Engineering (JTEC),* pp. 1-4, 2018.

[6] A. U. Aquino, M. E. M. Fernandez, A. P. Guzman, A. A. Matias, I. C. Valenzuela and E. P. Dadios, " An Artificial Neural Network (ANN) Model for the Cell Density Measurement of Spirulina (A. platensis)," *2018 IEEE 10th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment and Management (HNICEM),* pp. 1-5, 2018.

[7] P. J. M. Loresco, I. C. Valenzuela and E. P. Dadios, "Color space analysis using KNN for lettuce crop stages identification in smart farm setup," *TENCON 2018-2018 IEEE Region 10 Conference ,* pp. 2040-2044, 2018.

[8] F. Y. Osisanwo, J. E. T. Akinsola, O. Awodele, J. O. Hinmikaiye, O. Olakanmi and J. Akinjobi, "Supervised machine learning algorithms: classification and comparison," *International Journal of Computer Trends and Technology (IJCTT),* vol. 3, no. 48, pp. 128-138, September 2017.

[9] H. Uysal, A Genetic Programming Approach to Classification Problems, GRIN Verlag, 2016.

[10] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proceedings of COMPSTAT'2010*, Physica-Verlag HD, 2010.

[11] V. . Bewick, L. . Cheek and J. . Ball, "Statistics review 14: Logistic regression," *Critical Care,* vol. 9, no. 1, pp. 112-118, 2005.

[12] Z. . Minchen, W. . Weizhi, L. . Binghan and H. . Jingshan, "The Sigmoid function, where it is clearly demonstrated that the critical value range of is [−5, 5].," *PLOS ONE,* vol. , no. , p. , 2013.

[13] G. P.-S. C. M. W.J. Frawley, "Knowledge discovery in databases: An overview," *Knowledge Discovery in Databases, ,* pp. 1-27, 1991.

[14] L. Robinson and M. Smith, "Social Media and Mental Health - HelpGuide.org," 2020.

[15] B.-G. . Hu and W. . Dong, "A study on cost behaviors of binary classification measures in class-imbalanced problems," *arXiv: Learning,* vol. , no. , p. , 2014.

[16] M. Carbonero-Ruz, F. J. Martínez-Estudillo, F. Fernández-Navarro, D. Becerra-Alonso and A. C. Martínez-Estudillo, "A two dimensional accuracy-based measure for classification performance," *Information Sciences,* vol. 382, no. , pp. 60-80, 2017.

[17] I. . Visentini, L. . Snidaro and G. L. Foresti, "Diversity-aware classifier ensemble selection via f-score," *Information Fusion,* vol. 28, no. , pp. 24-43, 2016.

[18] I. Rish, "An empirical study of the naive Bayes classifier," 2001.

[19] J. Shen, S. Zhao, Y. Yao, Y. Wang and L. Feng, "A novel depression detection method based on pervasive EEG and EEG splitting criterion," *Proceedings - 2017 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2017,* Vols. 2017-Janua, pp. 1879-1886, 2017.

[20] D. e. a. Mittal, "An Effective Hybridized Classifier for Breast Cancer Diagnosis 2015," *IEEE International Conference on Advanced Intelligent Mechatronics (AIM),* 2015.

[21] P. e. a. Peduzzi, "A Simulation Study of the Number of Events per Variable in Logistic Regression Analysis," *Journal of Clinical Epidemiology,* vol. 49, p. 1373–1379, 1996.

[22] A. Shelar and C.-y. Huang, "Analyzing relationship: twitter tweet frequency with the stock prices of telecom companies," *Journal of Computing Sciences in Colleges,* vol. 34, no. 3, pp. 129-129, 2019.

[23] W. Gordon, "Understanding OAuth: What Happens When You Log Into a Site with Google, Twitter, or Facebook," 2020.

[24] A. . PappuRajan and S. P. Victor, "Web Sentiment Analysis for Scoring Positive or Negative Words using Tweeter Data," *International Journal of Computer Applications,* vol. 96, no. 6, pp. 33-37, 2014.

[25] T. . Singh and M. . Kumari, "Role of Text Pre-processing in Twitter Sentiment Analysis," *Procedia Computer Science,* vol. 89, no. , pp. 549-554, 2016.

[26] J. Yi, T. Nasukawa, R. Bunescu and W. Niblack, "Sentiment analyzer: extracting sentiments about a given topic using natural language processing techniques," *Proceeding of ICDM-03, the 3ird IEEE International Conference on Data Mining,* pp. 427-434, 2003.

[27] S. M. Kamruzzaman and C. M. Rahman, "Text Categorization using Association Rule and Naive Bayes Classifier," *arXiv: Information Retrieval,* vol. , no. , p. , 2010.

[28] G. M. Fung, O. L. Mangasarian and J. W. Shavlik, "Knowledge-Based Support Vector Machine Classifiers," *Advances in neural information processing systems,* pp. 537-544, 2002.

[29] F. Kabir, S. A. Siddique, M. R. A. Kotwal and M. N. Huda, "Bangla text document categorization using Stochastic Gradient Descent (SGD) classifier," *2015 International Conference on Cognitive Computing and Information Processing (CCIP),* pp. 1-4, 2015.

[30] P. Xu, F. Davoine and T. Denoeux, "Evidential Logistic Regression for Binary SVM Classifier Calibration," *In International Conference on Belief Functions,* pp. 49-57, 2014.