iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** XII **Month of publication:** December 2022

DOI: https://doi.org/10.22214/ijraset.2022.47955

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue XII Dec 2022- Available at www.ijraset.com

Statistical Tests for Comparing Machine Learning Algorithms

Koustubh Soman[1], Chinmay Kokate[2], Atharva Mohite[3], Ayush Vispute[4], Omkar More[5], Prof. Zarinabegam Kasimali Mundargi[6]
[1, 2, 3, 4, 5, 6]*Department of Artificial Intelligence and Data Science, Vishwakarma Institute of Technology, Pune*

Abstract: ***The average result of machine learning models is determined using naive 10-fold cross-validation, paired Student's t-test, McNemar's test. The algorithm with the best average performance should outperform those with the worst. But what if the difference in average results is due to a statistical anomaly? A statistical hypothesis test is used to determine whether the mean results of the differences between the three algorithms are real. Using statistical hypothesis testing, this study will show how to compare machine learning algorithms. When choosing a model, the output of several machine learning algorithms or simulation channels is compared. The model that performs best based on your performance measure becomes the final model that can be used to predict new data. With classification and regression prediction models, this can be done using traditional machine learning and deep learning methods. It is difficult to determine whether the difference between the two models is accurate or not.***

Keywords: Machine Learning, Classifiers, Data Mining Techniques, Data Analysis, Learning Algorithms, Supervised Machine Learning.

I. INTRODUCTION

Data can be understood by applying the specified structure to the result and then using statistical techniques to verify or invalidate the estimate. An estimate is called a hypothesis and validation done using statistical tests is called statistical hypothesis tests. If we want to make a claim about the distribution of the data or if the grouped results differ in applied machine learning, statistical hypothesis testing needs to be done. [1][2]. The data itself is uninteresting. The most interesting thing is how the data is interpreted. When it comes time to query the data and understand the discovery, then statistical methodologies will be used to provide certainty or probability about the answers. This kind of procedure is called significance testing or hypothesis testing [3]. The term theory might evoke the concept of scientific investigations in which a hypothesis is tested. This is a good step in the right direction. A hypothesis test statistically calculates a number for a given estimate. The results of the test allow the researcher to determine whether the estimate is valid or falsified. In particular, two examples that are widely used in machine learning are as follows: • A hypothesis test performed on data that followed a normal distribution. • Performed a test of the assumption that two samples are selected from the same population distribution. A null hypothesis, commonly known as hypothesis 0, is a statistical test hypothesis (H0 for short). The default assumption, often known as the "nothing has changed" assumption, is widely used. Because all available information indicates that the evidence shows that H0 can be rejected, the first hypothesis, often known as Hypothesis 1 or H1, is a violation of the assumption of the test. H1 is basically just shorthand for some other hypothesis. Hypothesis 0 (H0): The hypothesis of the test is correct and is rejected at a significant level. Hypothesis 1 (H1): At the given significance level, the test assumption fails and is rejected. Before he can reject or fail to reject the null hypothesis, the results of the test must be evaluated [3]. Regardless of the level of significance, testing the results of hypotheses may contain errors. By estimating a specified structure for the outcome and then using statistical techniques to verify or invalidate the estimate, the data can be understood. An assumption is called a hypothesis, and the statistical tests used to test it are called statistical hypothesis tests. Whenever there is a need to examine the distribution of data or when group results differ [2-4].

Interpreting the P-value: What is the significance of the p-value? It is possible to determine whether a result is statistically significant by calculating the p-value. For example, it can perform a normality test on sample data and find that the sample data rarely differs from a Gaussian distribution, thereby rejecting the null hypothesis [5]. A hypothesis test in statistics returns a p-value as the result. This is a statistic that can be used to evaluate or quantify the results of a test and decide whether to reject the null hypothesis. This is achieved by comparing the p-value with the significance level, which is a predetermined value [5]. Alpha is widely used to indicate the degree of significance. The most commonly used alpha value is 0.05. A smaller alpha value, such as 0.01 percent, indicates a more robust interpretation of the null hypothesis [6].

The previously obtained alpha value will be compared with the p-value. The statistical significance of the result is obtained when the alpha is greater than the p-value. This means that there has been a change: The initial hypothesis is now rejected [7] [8]. The null hypothesis is not rejected if the alpha is less than the p-value, which means that the result is not significant. The null hypothesis must be rejected if alpha is equal to the p-value, which means it is a significant result [9]. For example, if a test was run to see if a sample of data was normal and the p-value was 0.07, one might say the following: The test failed to reject the null hypothesis at the 0.05 significance level, indicating that the sample of data was normal. . You can generate a confidence level for a hypothesis based on observed sample data by subtracting 1 from the significance level [10].

II. LITERATURE REVIEW

Machine learning algorithms use a variety of statistical, probabilistic, and optimization methods to learn from past experience and uncover useful patterns from large, unstructured, and complex data sets [1]. These algorithms have a wide range of applications, including automatic text categorization [2], network intrusion detection [3], spam filtering [4], credit card fraud detection [5], customer purchasing behavior detection [6], manufacturing process optimization [7] and disease modeling [8]. Most of these applications have been implemented using supervised variants [4, 5, 8] of machine learning algorithms rather than unsupervised ones. In the supervised variant, a prediction model is developed by learning a dataset where the label is known, and accordingly the outcome of unlabeled examples can be predicted [9].

The subject of this research is primarily the analysis of the performance of disease prediction approaches using various variants of supervised machine learning algorithms. Disease prediction and, in a broader context, medical informatics have received significant attention from the data science research community in recent years. The reason is primarily the wide adaptation of computer technologies in health care in various forms (e.g. electronic health records and administrative data) and the subsequent availability of large health databases for researchers. These electronic data are used in a wide range of areas of health care research, such as health care utilization analysis [10], measurement of hospital care network performance [11], examination of patterns and costs of care [12], development of disease risk prediction. model [13, 14], tracking chronic diseases [15] and comparing disease prevalence and treatment outcomes [16]. Our research focuses on disease risk prediction models involving machine learning algorithms (eg, support vector machine, logistic regression, and artificial neural network), specifically supervised learning algorithms. Models based on these algorithms use labeled patient training data for training [8, 17, 18]. For the test set, patients are classified into several groups such as low risk and high risk. With the growing applicability and effectiveness of supervised machine learning algorithms for predictive disease modeling, the scope of research seems to be advancing. Specifically, we found little research that comprehensively reviews published articles using various supervised learning algorithms for disease prediction. This research therefore aims to identify key trends between different types of supervised machine learning algorithms, their performance accuracy and the types of diseases studied. In addition, the advantages and limitations of various supervised machine learning algorithms are summarized. The results of this study will help researchers better understand the current trends and foci of disease prediction models using supervised machine learning algorithms and formulate their research goals accordingly. In comparing between different supervised machine learning algorithms, this study reviewed existing literature studies that used such algorithms for disease prediction according to the PRISMA guidelines [19]. More specifically, this article considered only those studies that used more than one supervised machine learning algorithm for a single disease prediction in the same research setting. This made the main contribution of this study (i.e. the comparison between different supervised machine learning algorithms) more accurate and comprehensive, as comparing the performance of one algorithm in different study settings can be biased and generate erroneous results [20]. Traditionally, standard statistical methods and the physician's intuition, knowledge and experience were used for prognosis and disease risk prediction. This practice often leads to unwanted distortions, errors and high costs and negatively affects the quality of services provided to patients [21]. With the increasing availability of electronic health data, more robust and advanced computational approaches such as machine learning have become more practical in disease prediction. In the literature, most of the related studies used one or more machine learning algorithms to predict a specific disease. For this reason, the primary objective of this study is to compare the performance of different supervised machine learning algorithms for disease prediction.

III. METHODOLOGY/EXPERIMENTAL

A. Use McNemar's test

McNemar's test can be used when we need to compare the performance of two classifiers when we have matched pairs. The test works well if there are many different predictions between two classifiers A and B, then if we have a lot of data. Using this test, we are able to compare the performance of two classifiers on N items with one set of tests, unlike how you did in a paired t-test.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue XII Dec 2022- Available at www.ijraset.com

1) Assumptions

a) Random Sample
b) Independence
c) Mutually exclusive groups

2) Contingency Table

It's a tabulation or count of two categorical variables. In case of the McNemar's test, we are interested in binary variables correct/incorrect or yes/no for a control and a treatment or two cases. This is called a 2×2 contingency table.
The contingency table may not be intuitive at first glance. Let's make it concrete with a worked example.
Consider that we have two trained classifiers. Every classifier makes binary class prediction for each of the 10 examples in a test dataset. The predictions are evaluated and determined to be correct or incorrect.

B. Paired t-test

1) Take a sample of N observations (obtained from k-fold cv). These results are assumed to come from a normal distribution with a fixed mean and variance.
2) Calculate the sample mean and sample variance for these observations.
3) Calculate the t-statistic.
4) Use a t-distribution with N-1 degrees of freedom to estimate how likely it is that the true mean is within a given range.
5) Reject the null hypothesis at the p significance level if the t-statistic does not lie in the following interval:

$$\text{INTERVAL}: \left[-t_{p/2,n-1}, +t_{p/2,n-1}\right]$$

Fig.1 Interval

C. k-fold Cross Validation

Cross-validation is a resampling procedure used to evaluate machine learning models on a limited sample of data.
The procedure has a single parameter called k, which refers to the number of groups into which the given data sample is to be divided.
As such, the procedure is often called k-fold cross-validation. When a specific value for k is chosen, it can be used in place of k in the model reference, such as k=10 becomes 10-fold cross-validation.
Cross-validation is primarily used in applied machine learning to estimate the skill of a machine learning model on unseen data. This means using a limited sample to estimate how the model is expected to perform in general when used to make predictions on data not used during model training.
It is a popular method because it is easy to understand and because it generally results in a less biased or less optimistic estimate of model skill than other methods such as a simple train-test split.

The general procedure is as follows:

1) Shuffle the dataset randomly.
2) Split the dataset into k groups
3) For each unique group:
4) Take the group as a hold out or test data set
5) Take the remaining groups as a training data set
6) Fit a model on the training set and evaluate it on the test set
7) Retain the evaluation score and discard the model
8) Summarize the skill of the model using the sample of model evaluation scores

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue XII Dec 2022- Available at www.ijraset.com

D. Machine Learning Models

1) Linear Regression: LR assigns a theta weight parameter to each of the training functions. The predicted output (h(θ)) will be a linear function of the properties and coefficients θ.

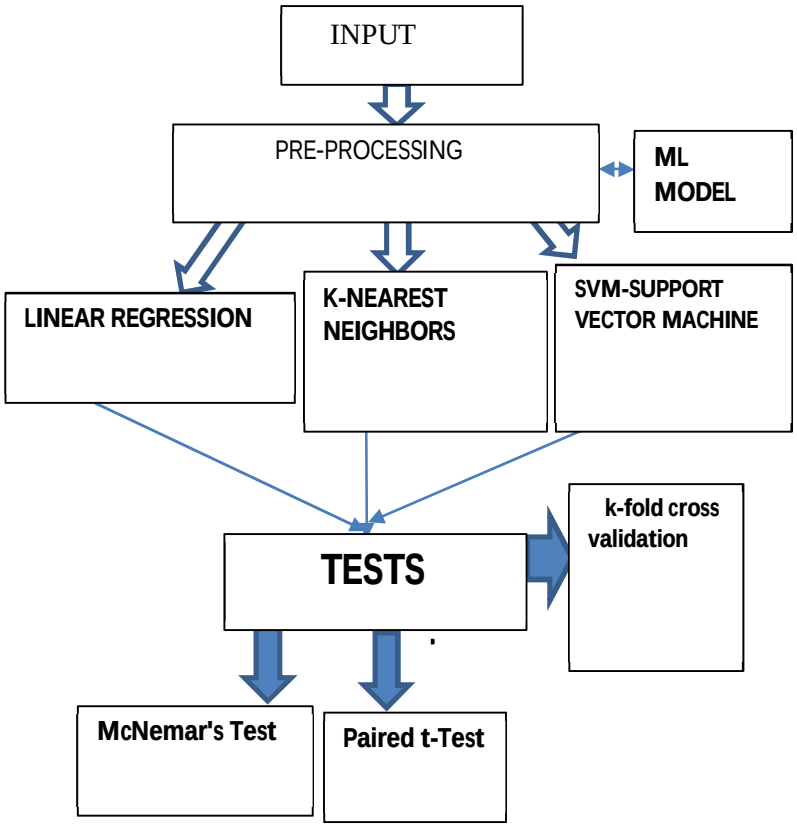
$$h_\theta = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ldots$$

Fig.2 linear regression output

During the beginning of training, each theta is randomly initialized. However, during training, we correct the theta corresponding to each element so that the loss (a metric of the deviation between expected and predicted output) is minimized. A gradient descent algorithm will be used to align the θ values in the correct direction. In the diagram below, each red dot represents the training data and the blue line shows the derived solution.

2) K-nearest neighbors: K-nearest neighbors is a non-parametric method used for classification and regression. It is one of the simplest ML techniques used. It is a lazy learning model with local approximation. Basic theory: The basic logic of KNN is to examine your surroundings, assume that the test data point is like them, and infer the output. At KNN, we look for neighbors and come up with a forecast. In the case of KNN classification, majority voting is used for the k closest data points, while in KNN regression, the average of the k closest data points is calculated as the output. As a general rule, we choose odd numbers like k. KNN is a lazy learning model where computations only happen at runtime.

3) Support Vector Machine: Support Vector Machine or SVM is one of the most popular supervised learning algorithms used for both classification and regression problems. However, it is primarily used for classification problems in machine learning. The goal of the SVM algorithm is to create the best line or decision boundary that can segregate the n-dimensional space into classes so that we can easily place a new data point into the correct category in the future. This best decision boundary is called the hyperplane. SVM selects extreme points/vectors that help in creating the hyperplane. These extreme cases are called support vectors, and thus the algorithm is called a support vector machine.



IV. RESULTS AND DISCUSSIONS

A. *LR vs SVM*

SVM supports both linear and non-linear solutions using the kernel trick. SVM handles outliers better than LR. Both work well when the training data is less and there are a large number of features.

B. *LR vs KNN*

KNN is a non-parametric model while LR is a parametric model. KNN is slow in real time because it has to track all the training data and find neighboring nodes, while LR can easily extract the output from the tuned θ coefficients.

C. *KNN vs linear regression*

KNN is better than linear regression when the data has high SNR.

D. *KNN vs SVM*

SVM takes care of outliers better than KNN. If the training data is much larger than not. functions (m>>n), KNN is better than SVM. SVM outperforms KNN when large features and less training data are available.

V. CONCLUSION

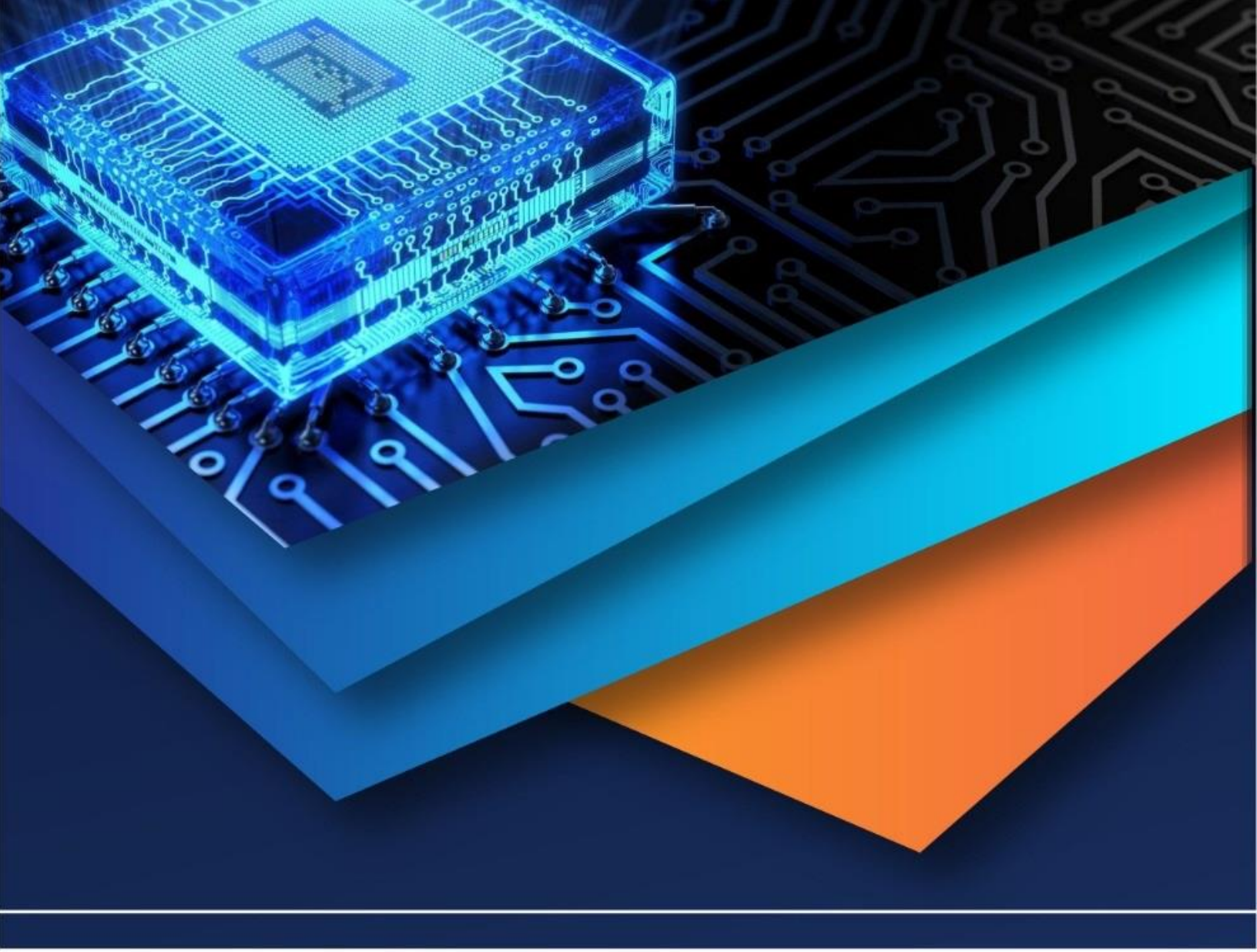
Hypothesis testing provides a reliable framework for making decisions about data about the population of interest. It helps the researcher to successfully extrapolate data from a sample to a larger population. Comparing the results you get once on different models to choose which one is the best is never a good method. Statistical tests allow us to objectively state whether one model works better.

This study demonstrates the use of statistical hypothesis tests to evaluate machine learning algorithms. In addition, it instructs researchers how to select models based on model performance averages, which can be misleading. A suitable methodology for comparing machine learning algorithms is five rounds of two-fold cross-validation using a modified Student's t-test. Compare algorithms using MLX tend machine learning and statistical hypothesis testing.

REFERENCES

[1] Komisi Penyiaran Indonesia, "Survei Indeks Kualitas Program Siaran Televisi Periode 5 tahun 2016," 2016.
[2] A. Abdallah, N. P. Rana, Y. K. Dwivedi, and R. Algharabat, "Social media in marketing : A review and analysis of the existing literature," Telemat. Informatics, vol. 34, no. 7, pp. 1177–1190, 2017. https://doi.org/10.1016/j.tele.2017.05.008
[3] W. G. Mangold and D. J. Faulds, "Social media: The new hybrid element of the promotion mix," Business Horizons., vol. 52, no. 4, pp. 357–365, 2009. https://doi.org/10.1016/j.bushor.2009.03.002
[4] M. H. Saragih and A. S. Girsang, "Sentiment Analysis of Customer Engagement on Social Media in Transport Online," in 2017 International Conference on Sustainable Information Engineering and Technology (SIET), 2017. https://doi.org/10.1109/SIET.2017.8304103
[5] J. H. Kietzmann, K. Hermkens, I. P. McCarthy, and B. S. Silvestre, "Social media? Get serious! Understanding the functional building blocks of social media," Business Horizons., vol. 54, no. 3, pp. 241–251, 2011. https://doi.org/10.1016/j.bushor.2011.01.005
[6] J. A. Morente-Molinera, G. Kou, C. Pang, F. J. Cabrerizo, and E. Herrera-Viedma, "An automatic procedure to create fuzzy ontologies from users' opinions using sentiment analysis procedures and multi-granular fuzzy linguistic modelling methods," Information Sciences, vol. 476, pp. 222–238, 2019. https://doi.org/10.1016/j.ins.2018.10.022
[7] Y. Lu, F. Wang, and R. Maciejewski, "Business Intelligence from Social Media: A Study from the VAST Box Office Challenge," Comput. Graph. Appl. IEEE, vol. 34 no 5, pp. 58-69, 2014. https://doi.org/10.1109/MCG.2014.61
[8] M. Yulianto, A. S. Girsang, and R. Y. Rumagit, "Business Intelligence for Social Media Interaction In The Travel Industry In Indonesia," J. Intell. Stud. Bus., vol. 8, no. 2, pp. 72–79, 2018.
[9] H. H. Do, P. W. C. Prasad, A. Maag, and A. Alsadoon, "Deep Learning for Aspect-Based Sentiment Analysis: A Comparative Review," Expert Syst. Appl., vol. 118, pp. 272–299, 2019. https://doi.org/10.1016/j.eswa.2018.10.003
[10] Y. Fang, H. Tan, and J. Zhang, "Multi-strategy sentiment analysis of consumer reviews based on semantic fuzziness," IEEE Access, vol. 6, no. c, pp. 20625–20631, 2018. https://doi.org/10.1109/ACCESS.2018.2820025
[11] H. Isah, P. Trundle, and D. Neagu, "Social Media Analysis for Product Safety using Text Mining and Sentiment Analysis," in 2014 14th UK Workshop on Computational Intelligence (UKCI), 2014. https://doi.org/10.1109/UKCI.2014.6930158
[12] P. Ducange, M. Fazzolari, M. Petrocchi, and M. Vecchio, "An effective Decision Support System for social media listening based on cross-source sentiment analysis models," Eng. Appl. Artif. Intell., vol. 78, no. October 2018, pp. 71–85, 2019. https://doi.org/10.1016/j.engappai.2018.10.014
[13] M. S. Omar, A. Njeru, S. Paracha, M. Wannous, and S. Yi, "Mining Tweets for Education Reforms," in 2017 International Conference on Applied System Innovation (ICASI), 2017. https://doi.org/10.1109/ICASI.2017.7988441
[14] P. F. Kurnia and Suharjito, "Business Intelligence Model to Analyze Social Media Information," Procedia Comput. Sci., vol. 135, no. September, pp. 5–14, 2018.https://doi.org/10.1016/j.procs.2018.08.144

[15] P. F. Kurnia, "Perancangan dan implementasi bisnis intelligence pada sistem social media monitoring and analysis (studi kasus di pt. dynamo media network)," Bina Nusantara University, 2017.
[16] M. A.- Amin, M. S. Islam, and S. Das Uzzal, "Sentiment Analysis of Bengali Comments With Word2Vec and Sentiment Information of Words," in 2017 International Conference on Electrical, Computer and Communication Engineering (ECCE), 2017. https://doi.org/10.1109/ECACE.2017.7912903
[17] R. Kimball and M. Ross, The Kimball Group Reader: Reader: Relentlessly Practical Tools for Data Warehousing and Business Intelligence. 2010.
[18] C. Vercellis, Business Intelligence, Data Mining and Optimization for Decision Making. John Wiley & Sons, Ltd, 2009.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY