A Comparative Study of Some Selected Classifiers on an Imbalanced Dataset for Sentiment Analysis

Mohammed Ali Kawo[1]; Dr. Garba Muhammad[2]; Dr.Danlami Gabi[3] and Dr. Musa Sule Argungu[4]
[1]Department of Computer Science, Federal University Gusau, Zamfara State, Nigeria
[1, 2, 3,4] Department of Computer Science, Kebbi State University of Science and Technology, Aliero, Nigeria

Abstract:- Extracting subjective data from online user generated text documents is made quite easy with the use of sentiment analysis. For a classification task different individual algorithms are applied to a review dataset in which most classifiers produce accurate results while others produce limited and inaccurate predictions. This research is to evaluate various machine learning algorithms for online dataset classification, where same set of data will be used to test four different machine learning algorithms: Naive Bayes, Support Vector machine, K-nearest neighbor and Decision tree. In order to determine which machine learning model will perform best in sentiment analysis as a constant issue. In this research, our primary goal is to identify the most effective machine learning model for sentiment analysis of English texts among the aforementioned classifiers. Their robustness will be tested and classified with an imbalanced dataset Kaggle.com a Machine learning repository. The dataset will first undergo data preprocessing in order to enable analysis, and then feature extraction for the base classifiers performance and accuracy which will be carried out in Jupyter notebook from Anaconda. Each machine learning algorithm performance scores will be calculated for higher accuracy using confusion matrix, F1-score, precision and recall respectively.

Keywords:- *Machine Learning Algorithms, Sentiment Analysis, Imbalanced, Confusion Matrix.*

I. INTRODUCTION

Machine learning is the concept of self-learning (George & Srividhya, 2022). It is a subset of artificial intelligence which involves training of computer to learn and improve from data without being thoroughly or detailed programmed. It deeply relied on algorithms and statistical models to recognize patterns and make predictions and decisions based on the input data. Machine leaning processes large amount of data to discover insights and develop automated responses and actions, enabling computers to perform task and improve their performance overtime (George & Srividhya, 2022).

Sentiment analysis examines how individuals express their ideas, sentiments, assessments, attitudes, and emotions in written language (Kumar et al., 2023). The inspection of views or feelings from text data is known as sentiment analysis or opinion mining. Sentiment analysis determines a person's opinion or sentiment toward a particular incident (Kawade & Oza, 2017). In order to perform sentiment analysis, we must provide a text or document that can be examined and that can provide a system or model that summarizes the opinions expressed in the text (Krishna, 2020). Customer's sentiment about company's goods and services is determined by comments and reviews from other users, it has proven extremely helpful in practically every business and social arena (Kumar et al., 2023). Sentiment analysis involves a variety of techniques which includes Natural Language Processing (NLP), Machine Learning (ML), Deep Learning (DL), Ensemble Methods and Hybrid Techniques.

(Kasthuri & Jebaseeli, 2020) Many studies concentrated on using standard classifiers to handle most problems such as the maximum entropy, naive Bayes, decision tree, K-nearest neighbor and support vector machine. But in order to improve the classification accuracy on sentiment analysis a substantial and robust classifier must to be obtained.

As a text classifier that can categorize text into different sentiments, sentiment analysis also known as opinion mining is useful for reviews of movies, products, customer services, opinions about any event, such as politics, societal activities (Kawade & Oza, 2017). Sentiment analysis is also useful for identifying people's opinions about any event like academics, practitioners and in human computer interaction, as well as those in other disciplines like sociology, marketing, economics and advertising (Bahwari, 2019). It can also be used to determine whether a particular item or service is good or bad, preferred or not preferred, and polarity of text (positive, negative, or neutral).

Due to the recent rapid rise of social platforms, a great deal of research in the field of sentiment analysis has focused on social medias. In order to improve company or find solutions to a variety of real world issues, practitioners and researchers have been working tirelessly to investigate and analyze this huge amount of data. They have done this by utilizing the daily interactions and ever growing user generated material that the websites facilitate (Agustini, 2021).

Educators must comprehend the views and feelings of their students, just as organizations must comprehend the thoughts of their clients. In an educational setting, sentiment analysis is also very useful where teachers and students are the driving forces behind the advancement of every nation's educational infrastructure (Alade & Nwankpa, 2022). In most cases, the creation of opinion mining or sentiment analysis systems in education is to find out what students think about education and how to improve the sector.

Sentiment analysis is the act of making assessments of people's ideas, imaginations, and personalities built on their written words, feelings, various picture types including emoticons, behavior, artwork, and other visual signs. Even though sentiment analysis is extensively used in so many domains, it still lack some areas where it application is needed and the best models that can effectively perform the analysis and predictions accurately is yet to be defined.

II. LITERATURE REVIEW

Sentiment analysis of review datasets using Naïve Bayes and K-NN Classifier as the two supervised methods used with two datasets namely film and hotel, (Bahwari, 2019) the more training data that is entered the better the accuracy obtained in the NB algorithm with the dataset film but for the K-NN method, accuracy is obtained randomly. (Tan et al., 2023) the authors set out to develop a sentiment analyzer that could accurately classify the polarity of text with outstanding precision. To do this, they employed five distinct machine learning techniques: Logistic Regression, Bernoulli Naive Bayes, Naive Bayes, and linear support vector classification where Naïve Bayes outperforms all other classifiers.

SVM is used to identify slogs. It was determined which models were most useful for logging web frameworks that used web indexes (Meenu, 2019). In (Meenu, 2019) authors suggested several grouping computations for Sequential Minimal Optimization (SMO), Logistic Regression, Decision Trees, Naïve Bayes, Classification, and Regression Trees to identify phishing mails in a coordinated manner across controlled and unsupervised methods.

In (Zishumba, 2019) Machine learning techniques such as Support Vector Machine, bag-of-words model, and Naïve Bayes are used for sentiment analysis of digital texts. In (Agustini, 2021) author employed a number of classifiers to assess a dataset of movie reviews and divide them into positive and negative categories. Out of 85,600 user comments, Logistic Regression performed the second best, with an accuracy rate of 99.46%. in another studies of (Agustini, 2021) author applied multiple classifiers to examine a dataset of movie reviews and classify them as favorable or unfavorable. With an accuracy of 99.46% for 85,600 user reviews, Logistic Regression delivered the second-best results.

(George & Srividhya, 2022) provides a successful method for creating precise classifiers for the Usenet2 dataset. The base classifiers used in the recommended approach are Naïve Bayes, Support Vector Machine, and Genetic Algorithm. In their work both homogeneous and heterogeneous models are constructed and classification accuracy improved significantly by the suggested ensemble bagged techniques compared to the base classifiers.

According to (Mostafa et al., 2021), Support Vector Machine, Bayesian, and Entropy classifiers were used to determine the sentiment polarity of tweets that yielding positive, negative and impartial tweets. These three distinct methods for classifying Twitter material according to phrases in supervised machine learning approaches were applied to trained datasets in three different ways. However, in order to obtain precise and trustworthy predictions many classifiers are combined using ensemble approaches.

(Kumar et al., 2023) conducts sentiment analysis on the Twitter140 dataset using Decision Tree, Logistic Regression, and Support Vector Machine. Within the biased techniques, these algorithms are very common. One of the struggles in machine learning sentiment analysis is the ability to acquire large amount of data for better classification (Lazrig & Humpherys, 2022).

III. BASE CLASSIFIERS FOR SENTIMENT ANALYSIS

Among the most innovative cutting edge technologies of the twenty-first century is predicted to be machine learning (Jordan & Mitchell, 2020). Despite the fact that the future cannot be predicted, society must start considering ways to optimize its advantages. To acquire more insight on our research, current and advanced reviews were explored in machine learning in other to establish more facts on the widely used machine learning algorithms to be used in our research, which are:

➢ *Naive Bayes:*

The Naive Bayes model can handle large amounts of data and is robust against complicated classification methods (George & Srividhya, 2022). Naïve Bayes theory is explained by the following equation: P(H|E) = (P(E|H) * P(H))/P(E). Where P(H|E) signify the prior probability of the hypothesis given that the evidence is true, P(E|H) is the likelihood of the evidence given that the hypothesis is true while P(H) is the prior probability of the hypothesis and P(E) is the prior probability that the evidence is true. (Patel, 2017) Predicting the correct class for a freshly produced instance and being simple to use are the main advantages of this classifier.

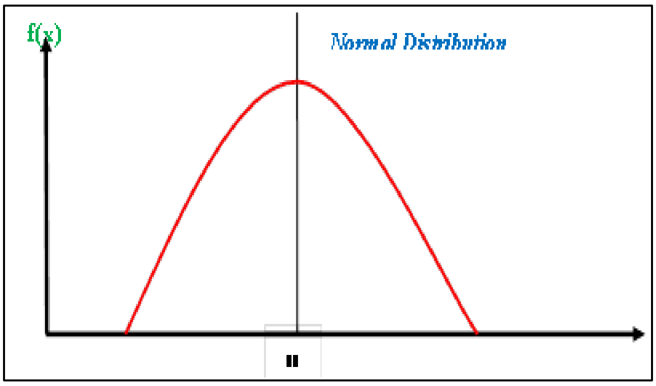

Fig 1 Naive Bayes (Gaussian Distribution)

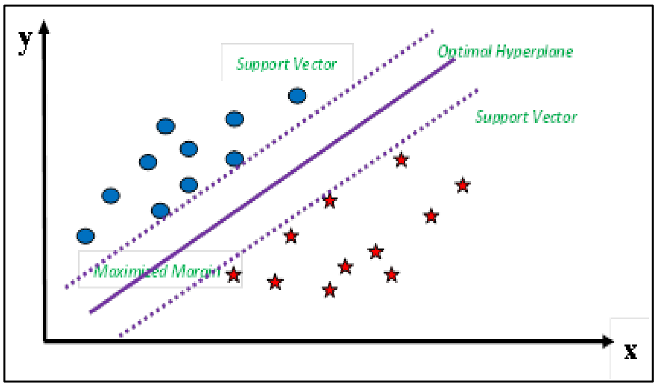

Fig 2 Support Vector Machine

- *Support Vector Machine:*

When it comes to nonlinear regression and classification tasks, support vector machines (SVMs) are essentially binary classifiers that work well at categorizing both linear and nonlinear data (Patel, 2017). SVMs handles overfitting problems that occur in high dimensional environments due to its global optimization base and it helpful in variety of applications (Liakos et al., 2018). The process presents each data point as a point in an n-dimensional space, where 'n' is the total number of features you possess. Each feature's value is represented by a unique coordinate. Finding which can be utilized to divide a certain class is the next stage in the classification process (George & Srividhya, 2022).

- *Decision Tree:*

Regression and classification may both be done using the decision tree due to its tree-like structure. Using decision trees, one can create a training model that can be used to predict the class or value of the destination variable. (Moret, 2019),The application of decision trees is very advantageous in many fields, such as databases, taxonomy and identification, machine diagnosis, switching theory, pattern recognition, decision table programming, and algorithm analysis. The diagrammatic representation of Decision Tree is illustrated in figure 3 below

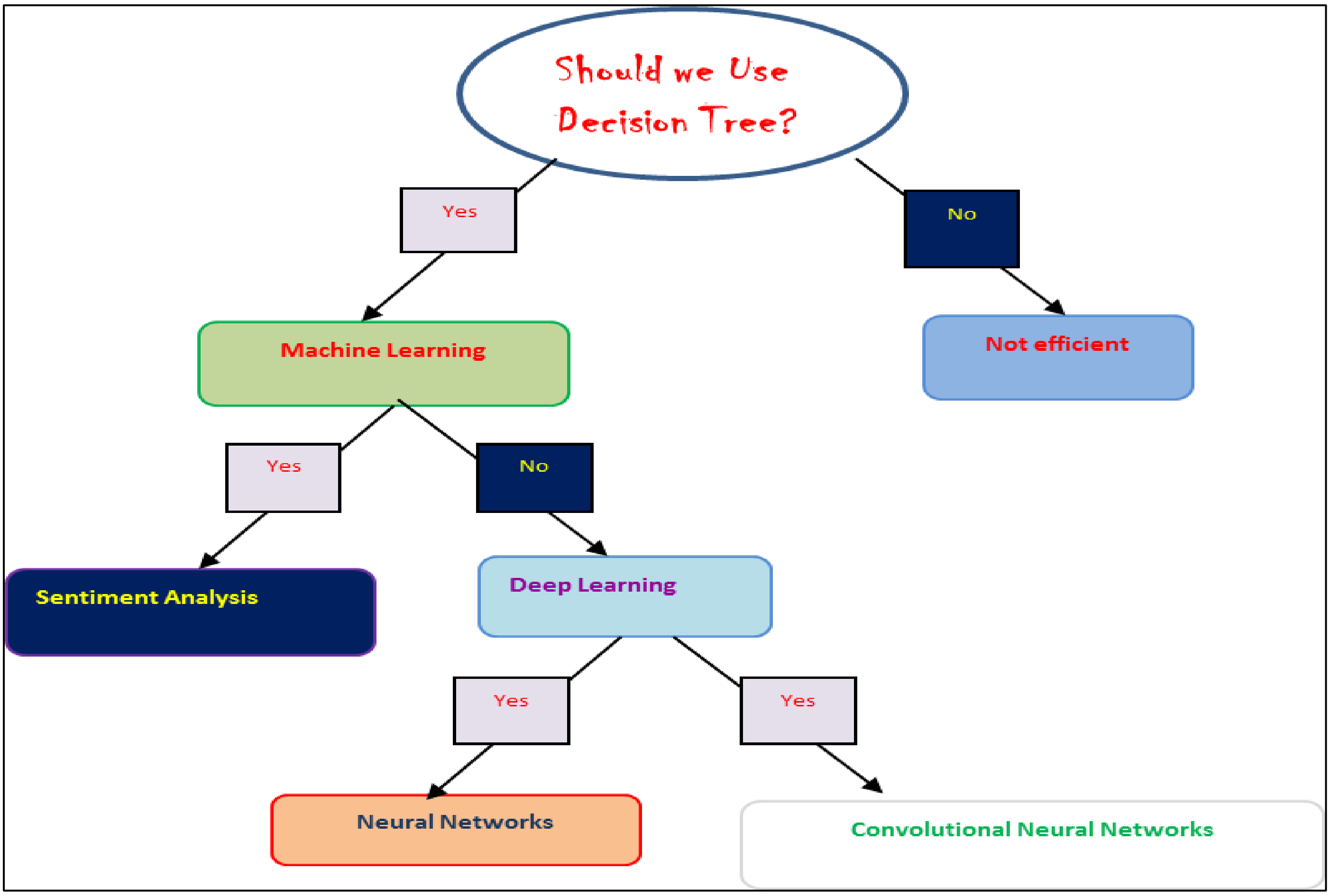

Fig 3 Decision Tree Diagram

- *K-Nearest Neighbors:*

One of the simplest machine learning algorithms and a theoretically valid method is the KNN technique, which was first proposed by Cover and Hart in 1967. The idea behind KNN is very simple and straightforward: given a sample, if the K closest neighbors (i.e., most similar samples) in the feature space are also samples in that class, then this sample also belongs to that class. The classification outcome of the sample is directly affected by the choice of K values (Feng et al., 2023). Figure 4 below displays how K-Nearest Neighbor took values of a different sample.

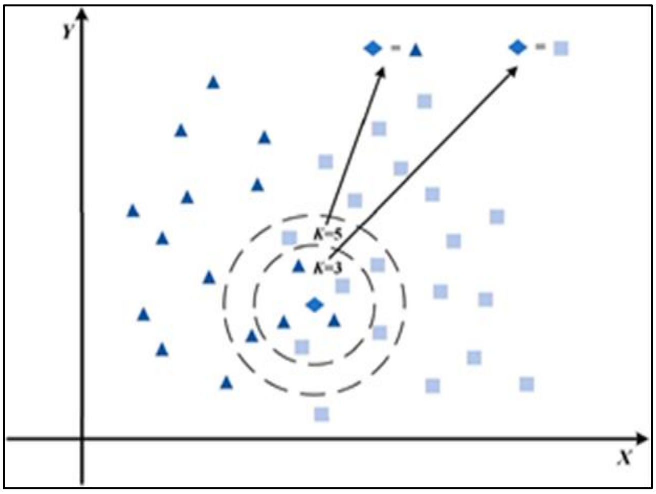

Fig 4 K-Nearest Neighbor

IV. DATASET

Datasets are very essential in data analysis and machine learning. A vigorous decision making process for organizations and the entire nation will be greatly enhanced by the meaningful utilization of data. In order to carry out the proposed research, a life and health insurance company imbalanced dataset from Kaggle.com a machine learning repository is chosen. This dataset is in 'csv' format that comprises of 267,507 records with fourteen input features.

For machine learning practitioners working on binary classification and sentiment analysis tasks frequently find imbalanced datasets as a barrier in detecting tasks such as fraud, spam, diseases and hardware faults. (George & Srividhya, 2022) A dataset that is imbalanced comprises of two distinct observations one belonging to the majority class and the other to the minority class. Figure 5 below shows the schema of imbalanced dataset.

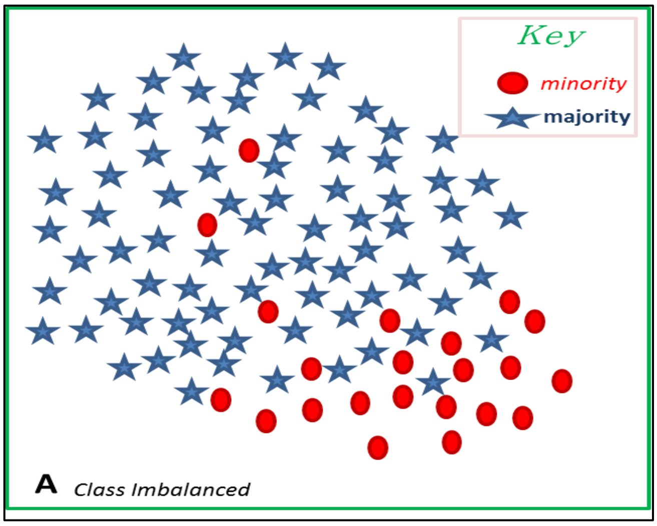

Fig 5 An Imbalanced Dataset

When most of the data in each class is evenly distributed then the majority of conventional data classification techniques can be implemented with proficiency in terms of total classification accuracy. However, when categorizing an imbalanced dataset that contained some examples from the interest group, these classifiers will unable to do any better (Thesis, 2023).

V. METHODOLOGY

The most essential segment of our research is enclosed in this section, where the techniques and algorithms that will be applied to the datasets in order to obtain the desired results. In our research, the aforementioned classifiers task would be to forecast using the provided input features of the imbalanced dataset, to determine whether the insured customers will be willing to sign up for vehicle insurance newly provided by the company.. All the four based classifiers will be utilized individually and their results will be compared in order to ascertain which classifier has the highest accuracy level. Figure 5 below illustrates our methodological workflow.

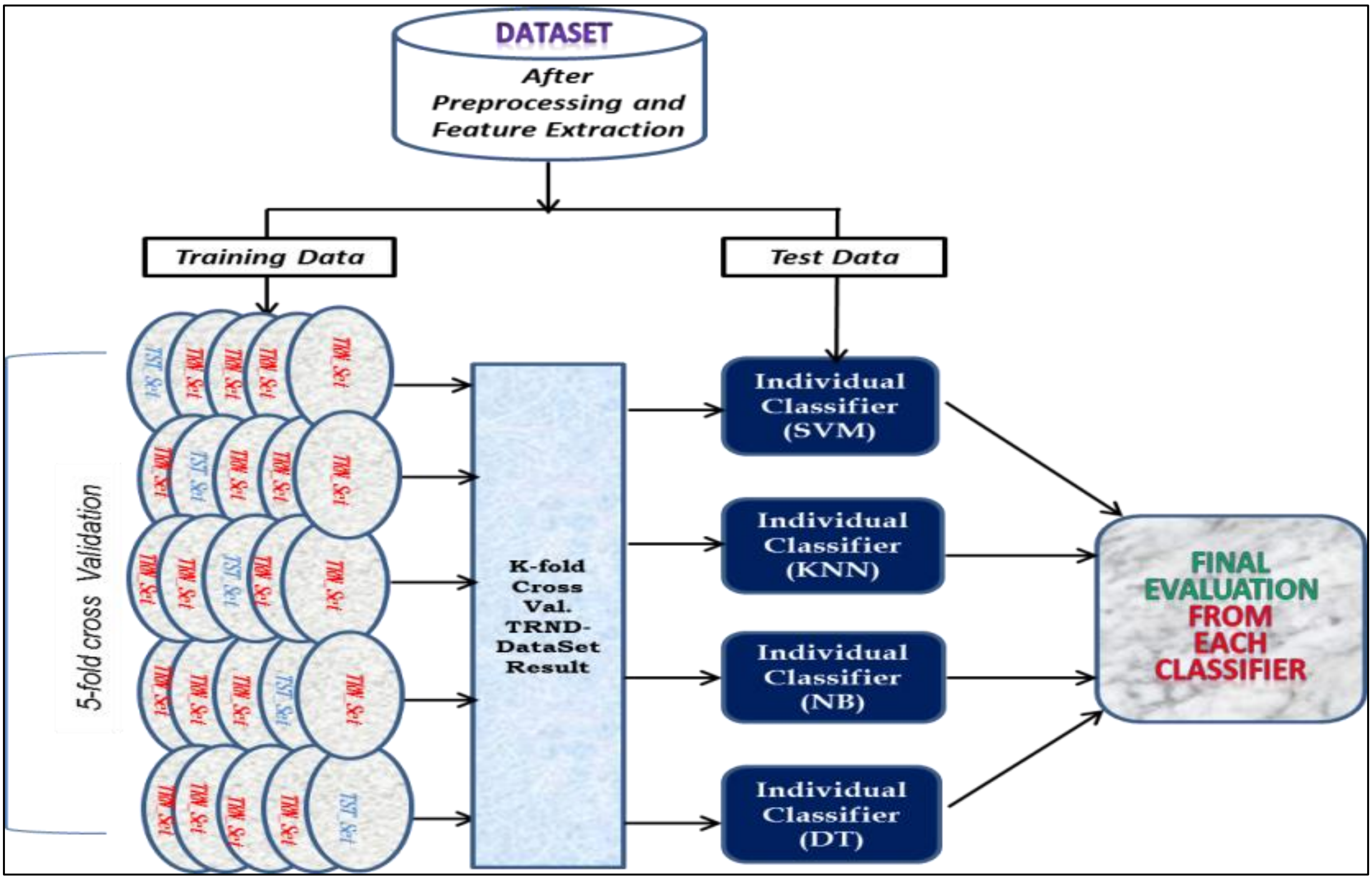

Fig 6 Proposed Workflow Diagram

➢ *Data Preprocessing*

Once data has been collected, data preparation mostly known as preprocessing is an essential step in sentiment analysis. It cleans, organizes, and scrubs raw data into a format that machine learning models can use for training and evaluation. Preprocessing also known as text filtering (Arya et al., 2019), is the process of removing noisy, unreliable and partial datasets through the use of tokenization, stemming, and vectorization techniques all of which are essential sub-steps in the process.

Scrubbing the data is a crucial first step in doing preprocessing for sentiment analysis. Scrubbing is the technical process of enhancing the dataset to increase its utility. This will require data that is redundant, incomplete, incorrectly formatted, or irrelevant to be edited and occasionally removed (Theobald, 2017).

➢ *Separating Training and Testing Data Set*

When using machine learning, we typically divide our original dataset into two subsets the training set and the testing set. We then fit our model using the train data in order to provide predictions for the test set. In order to divide the original datasets into training and testing sets since the dataset in an imbalanced one, we employed both stratified sampling techniques and k-fold cross-validation with k equal to 5. This will allow the sharing of similar representative samples from each class and will improve the quality and efficiency of the models to enable smooth comparison among them.

➢ *Training Model*

Ultimately, in our research, we employed four trending distinctive models such as K-nearest Neighbors (KNN), Decision trees (DT), Support Vector Machines (SVM) and Naïve Bayes (NB) to classify the dataset in order to ascertain which classifier will perform best in term of accuracy, precision, recall and f1-score respectively.

Our current research encountered a problem while using SVM classifier commonly known as support vector machine due to the enormous dataset, the SVM classifier took approximately two and half hours classifying dataset of about above three hundred thousand with just fourteen features. Consequent to that, we deployed DSVM classifier which is known as dual support vector machine due to the problem found. The DSVM is prompt in classifying large datasets with less time consumption.

➢ *Testing the Model*

Performance evaluation is a critical component of every research study. Given that it is essential to examine the behaviors of the system. A confusion matrix is used in machine learning to evaluate the effectiveness of a classification model (Zishumba, 2019). In case where the true values are known, it compares test result in tabular form. The performance of the suggested models in this research will also be evaluated using the confusion matrix.

VI. RESULT AND DISCUSSION

To the provided kaggle imbalanced dataset of life and health insurance Company, we implemented four classifiers such as Naïve Bayes, Support Vector Machine, K-Nearest Neighbor and Decision Tree, each with the same datasets that is split up into a training set and a testing set. The Jupyter notebook from Anaconda is used to explore the experiment which serves as the basis for the findings that are presented here. Various performance metrics including classification accuracy will be used to compare these classifiers; Precision, Recall, and F1-score values obtained from each classifier's will also serve as an additional means of assessing the accuracy on these classifiers.

Following the end of the training phase on each of these classifiers (NB, KNN, SVM, and DT), the dataset is applied to test the classification performances, table 1, below displays the confusion matrix of all the classifiers on stratify sampling technique that was used on the dataset in order to have equal representations in all the classes. While figure 6, below also shows the bar chart representation of all the four classifiers accuracy result.

Table 1 Performance Evaluation based on Confusion Matrix

Classifiers	Accuracy	Precision	Recall	F1-score
Naïve Bayes	77%	41%	84%	55%
K-N-Neighbor	80%	31%	16%	21%
SVM	84%	48%	10%	16%
Decision Tree	81%	43%	45%	44%

From table 1 above, the statistical values of all the four classifiers is show, where Support vector machine with the accuracy score of 84 percent outperform all other classifiers even though the recall and f1-score is low, while decision tree came next with 81 percent accuracy score with average scores in recall and f1 respectively.

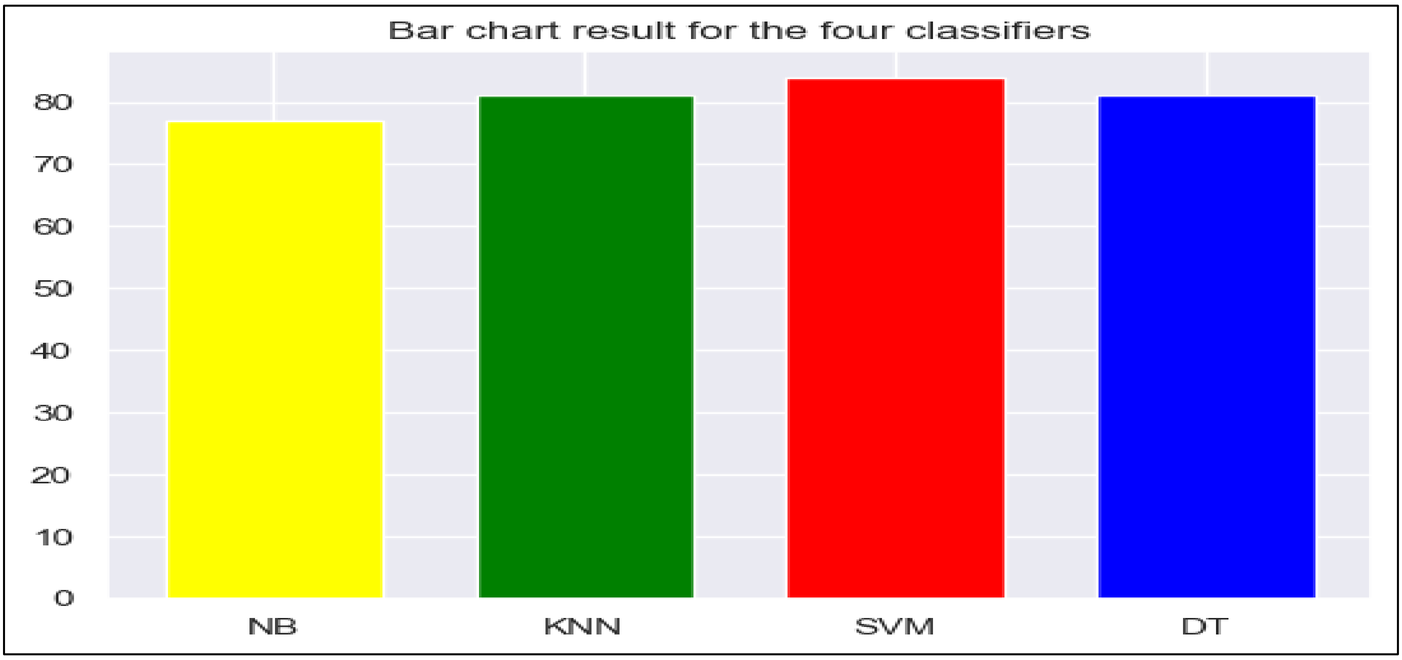

Fig 7 Bar Chart Accuracy Result for the Four Classifiers

Table 2 Performance Comparisons of Four Classifiers based on K-fold cross Validation

Classifiers	Accuracy K1	Accuracy K2	Accuracy K3	Accuracy K4	Accuracy K5
Naïve Bayes	0.76955	0.77583	0.77150	0.77039	0.77184
KNN	0.80439	0.80821	0.80617	0.80524	0.80613
SVM	0.83555	0.83713	0.83565	0.83674	0.83586
Decision Tree	0.81260	0.81791	0.81325	0.81303	0.81345

From all indication in table 2 above, it has been clearly shown that Support Vector Machine has the highest accuracy score of 84 percent followed by Decision Tree classifier with approximately 82 percent.

VII. CONCLUSION AND FUTURE WORKS

In summary, insurance companies, product companies, industries of all kinds, institutions and health practitioners can utilize machine learning method in analyzing sentiments. The life and health insurance company dataset used to predict if the customers are willing to apply for vehicle insurance in that same company. From the predicted analysis result in table one show the low score in precision which indicates low outcome of customers that are willing to review their vehicle insurance with the same company. Even though our target is to compare the classifiers, we still have to predict the outcome of the dataset. The frequent change in vocabularies and cultural diversities has raised a great challenge in the field of sentiment analysis. Every culture has a way of expressing emotions be it happiness or sadness. Contextual sensitivity is another factor that contributes to the challenges of sentiment analysis, since grammar continues to revolve every day. From the research carried out it has been proven that support vector machine has the highest classification score compared to naïve bayes, k-nearest neighbor and decision tree. This indicates that even in an imbalanced data classification process support vector machine still perform excellently. For the fact remains that support vector machine has the strength of handling complex regression or classification problems. In the future, deep learning should be compare with some of the base machine learning classifiers such as support vector machine in predicting sentiment analysis so as to standardize a model for sentiment analysis.

REFERENCES

[1]. Agustini, T. (2021). Sentiment Analysis on Social Media using Machine Learning-Based Approach. June, 544437.

[2]. Arya, P., Bhagat, A., & Nair, R. (2019). Improved Performance of Machine Learning Algorithms via Ensemble Learning Methods of Sentiment Analysis. 10(2), 110–116.

[3]. Bahwari. (2019). Sentiment Analysis Using Random Forest Algorithm - Online Social Media Based. Journal Of Information Technology AND ITS UTILIZATION, 2(2), 29–33. https://www.researchgate.net/publication/338548518_SENTIMENT_ANALYSIS_USING_RANDOM_FOREST_ALGORITHM_ONLINE_SOCIAL_MEDIA_BASED

[4]. Feng, W., Gou, J., Fan, Z., & Chen, X. (2023). An ensemble machine learning approach for classification tasks using feature generation. Connection Science, 35(1). https://doi.org/10.1080/09540091.2023.2231168

[5]. George, S., & Srividhya, V. (2022). Performance Evaluation of Sentiment Analysis on Balanced and Imbalanced Dataset Using Ensemble Approach. Indian Journal of Science and Technology, 15(17), 790–797. https://doi.org/10.17485/ijst/v15i17.2339

[6]. Ghosh, S., Hazra, A., & Raj, A. (2020). A Comparative Study of Different Classification Techniques for Sentiment Analysis. International Journal of Synthetic Emotions, 11(1), 49–57. https://doi.org/10.4018/ijse.20200101.oa

[7]. Jawale, S. (2019). Sentiment Analysis using Ensemble Learning. May.

[8]. Jordan, M. I., & Mitchell, T. M. (2020). Machine learning: Trends, perspectives, and prospects. Science, 349(6245), 255–260. https://doi.org/10.1126/science.aaa8415

[9]. Kawade, D. R., & Oza, D. K. S. (2017). Sentiment Analysis: Machine Learning Approach. International Journal of Engineering and Technology, 9(3), 2183–2186. https://doi.org/10.21817/ijet/2017/v9i3/1709030151

[10]. Kumar, S., Kaur, N., Kavita, & Joshi, A. (2023). Tweet sentiment analysis using logistic regression. July, 332–336. https://doi.org/10.1049/icp.2023.1801

[11]. Lazrig, I., & Humpherys, S. L. (2022). Using Machine Learning Sentiment Analysis to Evaluate Learning Impact. Information Systems Education Journal (ISEDJ), 20(1), 20. https://isedj.org/; https://iscap.info

[12]. Liakos, K. G., Busato, P., Moshou, D., Pearson, S., & Bochtis, D. (2018). Machine learning in agriculture: A review. Sensors (Switzerland), 18(8), 1–29. https://doi.org/10.3390/s18082674

[13]. Meenu, S. G. (2019). 154. Sunila. International Journal of Electronics Engineering (ISSN: 0973-7383, Volumne 11(• Issue 1), 965–970.

[14]. Mostafa, G., Ahmed, I., & Junayed, M. S. (2021). Investigation of Different Machine Learning Algorithms to Determine Human Sentiment Using Twitter Data. International Journal of Information Technology and Computer Science, 13(2), 38–48. https://doi.org/10.5815/ijitcs.2021.02.04

[15]. Patel, R. (2017). Sentiment Analysis on Twitter Data Using Machine Learning by Ravikumar Patel A thesis submitted in partial fulfillment of the requirements for the degree of MSc Computational Sciences The Faculty of Graduate Studies.

[16]. Tan, K. L., Lee, C. P., & Lim, K. M. (2023). A Survey of Sentiment Analysis: Approaches, Datasets, and Future Research. Applied Sciences (Switzerland), 13(7). https://doi.org/10.3390/app13074550

[17]. Theobald, O. (2017). Machine Learning For Absolute Beginners.

[18]. Zishumba, K. (2019). Sentiment Analysis Based on Social Media Data. Journal of Information and Telecommunication, 1–48. http://repository.aust.edu.ng/xmlui/bitstream/handle/123456789/4901/Kudzai Zishumba.pdf?sequence=1&isAllowed=y