IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 5, May 2022
Impact Factor: 6.252

Analysis of Different Machine Learning Algorithms Used for Spam E-mail Detection

Shubham Zanzad[1], Devansh Thard[2], Tushar Jarare[3], Santosh Shinde[4], Prof. B. S. Gayal[5]
Students, Dept. of Information Technology Engineering[1,2,3,4]
Guide, Dept. of Information Technology Engineering[5]
Sinhgad Academy of Engineering, Pune Maharashtra, India

Abstract: *From business to education, email is now used in almost every industry. Subcategories of email exist, such as ham and spam. Unsolicited email, also known as spam or junk email, is a sort of email that can be used to harm consumers by wasting their time, using up their computer resources, and collecting sensitive information. Every day, the amount of spam sent out increases alarmingly. For email and IoT service providers, spam detection and filtering have suddenly become substantial and pervasive concerns. Email filtration is one of the most essential and well-known advanced spam detection and prevention techniques. Many machine learning and deep learning algorithms have been used for this purpose, including Naive Bayes, decision trees, neural networks, and random forests. This article divides utility research approaches into applicable classifications based on machine learning tactics used in texting systems. The accuracy, precision, recall, and other performance characteristics of these approaches are all well assessed. Finally, broad ideas and prospective study directions are provided.*

Keywords: SVM, Decision Tree, K-Nearest Neighbor, Naïve Bayes, Boosting Algorithm

I. INTRODUCTION

Mail that really was spam the detection procedure begins with supervised message filtering and is followed by basic filtering strategies that can identify communications with specified characteristics. Digitised spam detection begins with the use of common utility survey approaches to develop spam detection models. Spam originates with unsolicited emails known as Unsolicited Bulk Email (UBE) or Unsolicited Industrial Email (UIE)from spammers and some verified institutes. Texting, on the other hand, is a completely cost-effective method for sending character messages to potential clients, with a higher response rate than spam coupled with e-mail and SMS, social networks like Twitter and Facebook, and instant messaging services like WhatsApp, among other contribute to the production of a big amount of spam on the Community In the absence of automated customs clearance at the moment of reception, spam detection is a time-consuming task. Rule-based filtration was an early classifier that allowed policies to be expressed more formally and applied across several client areas. It is made up of a series of predetermined criteria that are applied to an incoming message, and the message is tagged as spam if the checkpoint surpasses a certain level.

II. MOTIVATION

E-commerce companies have grown in popularity in recent years due to the wide range of services they provide and the convenience with which their products and reviews can be found online. Users have discovered that reading online reviews can help them make better judgments while shopping on these websites. Spammers have targeted etrade websites to review unwanted messages because of their popularity. On most e-commerce sites, customers can leave comments on products in a basic review section. Customers can offer feedback on their services and products on many review sites, including TripAdvisor.com, Zomato.com, Amazon.com, and Yelp.com. The term "user generated content" refers to online material. User-generated content (UGC) includes a plethora of interesting and useful information about products and services. Because there is no effective regulation of this content on the internet, scammers are encouraged to write false and misleading product information.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 5, May 2022
Impact Factor: 6.252

III. LITURETURE SURVEY

Ms. Sayali Kamble, Dr. S.M.Sangve "Real Time Detection of Drifted Twitter Spam Based on Statistical Features."[1] These activities allow people to share information and allow customers to discuss their conduct, revealing their reputation; they also serve as antecedents to various sorts of spam. Twitter's most popular subjects at any one time are leveraged to generate traffic and revenue. Spammers attempt to capture people's interest by sending tweets with irrelevant content, malicious links, and recurring topics. Because it's been a long time since someone sent an unpleasant tweet, there's a chance it'll be made public to suspects. As a result, it is critical to identify spam tweets as quickly as feasible. To reduce losses caused by unsolicited messages, real-time detection is required.

Thayakorn Dangkesee ,Sutheera Puntheeranurak "There are various systematic and standardized methods for spam detection that take the statistical features of tweets into account. URLs are assessed and tested for hazard using different APIs in the suggested extension. to detect spam on Twitter by using specific records" [2] As a result of the gift, Twitter's popularity is increasing. Customers who have tweeted can access data on Twitter. Meanwhile, here are some statistics produced by spammers who sincerely want to market their websites or services. They have an impact on ordinary users by taking advantage of consumer interests on Twitter channels, such as posting unwanted links and advertisements. To tackle spammers, many researchers have created anti-spam strategies. Recent research, on the other hand, has concentrated on how to create streaming spam detection systems. In this post, we proposed an adaptive record type for spam detection using a spam phrase collection and the company's URL-based protection solution. To analyse records with both all-inclusive and specific data types, we employed the Nave Bayes rule set. This can help to improve the overall performance of the spam detection. The application of our offered solutions will be demonstrated in the test results.

Rutuja Katpatal,Aparna Junnarkar "An Efficient Approach of Spam Detection in Twitter" [3] Spam on Twitter has become a serious problem in recent years. Late Paintings is especially interested in applying machine learning to detect spam on Twitter by analysing actual tweet components. Reduction is used to modify an existing machine by learning about the classifier. This is known as Twitter spam surfing. Lfun is a system that detects spam tweets that have been altered from untagged tweets and integrates them into a taxonomy with the sole objective of addressing this problem. Tweeting might assist you in locating spam messages. After a predetermined period of time, our recommended system will change the educational information, including deleting obsolete templates and clearing the region where extraneous data is stored.

Lekshmi M B,Deepthi V R, "Spam Detection Framework for Online ReviewsUsing Hadoop's Computational Capability"[4] Online ratings have become one of the most important factors for customers when shopping online. These profiles are used by teams and individuals to acquire the correct items and choose the right business. As a result, spammers and unethical dealers have created phoney reviews to promote their products to superior competitors. Spammers use sophisticated techniques to generate a flood of unsolicited mail ratings throughout the internet in hours. To address this issue, research has been conducted to develop effective methods for detecting spam reviews. Countless spam detection algorithms have been developed, the majority of which extract useful functions from textual content or employ gimmicky learning strategies. These methods pay minimal attention to the types of features that are extracted or the processing costs. NetSpam is a methodology for classifying a review dataset only on spam likelihood and mapping it to a method of identifying unsolicited mail that outperforms earlier work on accuracy forecast. An attribution technique on the evaluation dataset utilising the MapReduce function is proposed as a solution to improve the processing fee in this study. Hadoop makes use of parallel programming and MapReduce to process massive amounts of data. Parallelizing the NetSpam rule set and developing spam detection modes with improved prediction accuracy and processing load are part of the solution.

IV. SYSTEM ARCHITECTURE

The above graphic depicts the proposed device's architecture, and first and foremost, we'll educate the undesirable mail data set within the provided Architectural diagram, and then processing will be applied as the genuine global data made from faults. To obtain better results from a given data set, it is important to mine the records, and the information must be preprocessed before applying a classifier to the facts set. It consists of record cleansing, integration, and transformation and is critical. Before using any statistics mining approaches to acquire better results, five. normalise the complete statistics collection (normalisation is a system in which a database is based on a systematic manner of tables and results need to be in an unambiguous layout) to assess whether this is spam or not.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 5, May 2022
Impact Factor: 6.252

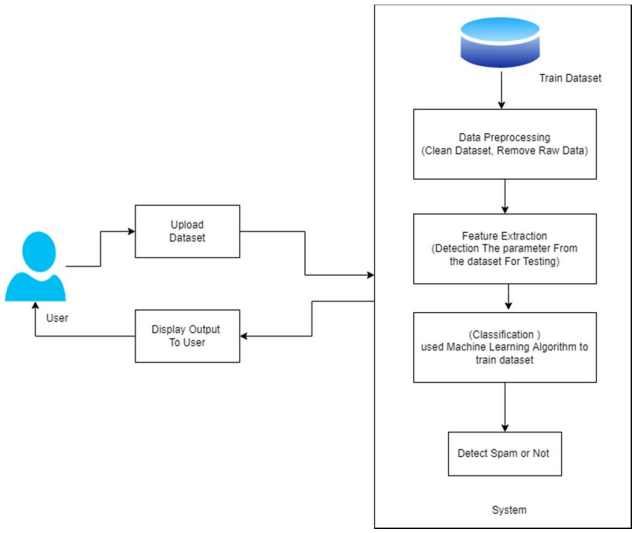

Figure 1: System Architecture

V. OBJECTIVE

The project's main goal is to lower the calculation time of the Machine and diverse datasets in order to compare the results of these methods against different testing sizes. To identify the most efficient algorithm and provide a user-friendly GUI. To determine the best classifier, compare results and learn how algorithms work. We will recommend the appropriate algorithm for your project.

VI. DATASETS USED

For this project we are considering 2 dataset which were open-source. There are many different datasets present as well but the data representation in the file is really messed. So after scrapping deep into the web we found out two datasets one from UCI Machine Learning and second one is from Kaggle. Analysis of these datasets is given below.

VII. DATA CLEANING

Data cleaning is an important step in building a machine learning model because it allows us to examine the dataset we're using, see if any of the items in the data have null values, rename columns for easier understanding, and remove duplicates and perhaps other columns which have no useful value for the project. Eradicating all of these values can make a significant difference in the machine learning model's performance. We are utilising a dataset from UCI Machine Learning for this project, which has a variety of columns. It is made up of a collection of items that are categorised as ham or spam communications. Also, this dataset has three columns that seemed pointless because they only included very few items below in a vast dataset, increasing the model's computation time. As a result, removing those columns significantly sped up the model's execution time. We've also changed the names of the remaining columns to 'target' and 'text.' The values in the target field have been converted to 0 for ham and 1 for spam. The text field holds the input text message that the dataset will process.

VIII. EXPLORATORY DATA ANALYSIS

To comprehend the data, we must undertake Exploratory Data Analysis whenever we perform any predictive modelling task. For this model, we only need to examine two columns in our dataset. So, after evaluating the data, we discovered that 87 percent of the values in the dataset are ham values and 13 percent of the values are spam, indicating that the data is imbalanced. Because the data for spam is now relatively limited, accuracy becomes a crucial metric for us to consider while analyzing the model.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 5, May 2022
Impact Factor: 6.252

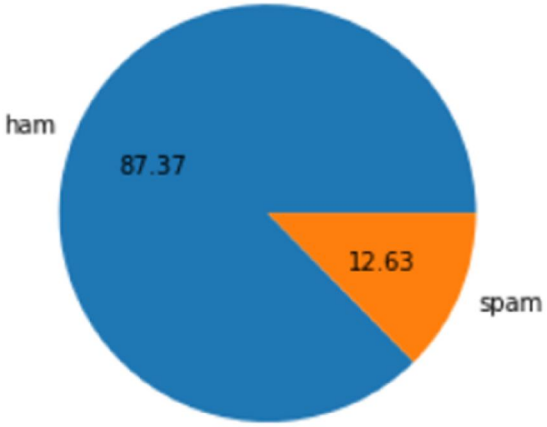

Fig 2. Dataset 1

For the second dataset, it appears to be more balanced; the percentage of spam and ham messages in the corpus is 44.28 percent and 55.72 percent, respectively, which is well balanced for an open source dataset. So after computing and evaluating accuracy and precision both will be equally important.

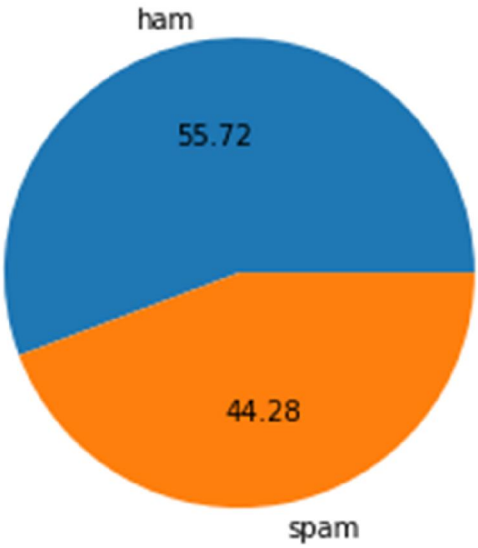

Fig 3. Dataset 2

WordCloud creation is also quite handy when it comes to examining the datasets that we use. The goal of WordCloud is to display on a canvas the text that is often used in the document or the input corpus. The greater the frequency, the larger the text size. So, after creating a WordCloud from these two datasets, we can better understand the dataset and utilise it as a parameter for filtering techniques in the future. The following are WordClouds for the two datasets, Spam and Ham Corpus:

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 5, May 2022
Impact Factor: 6.252

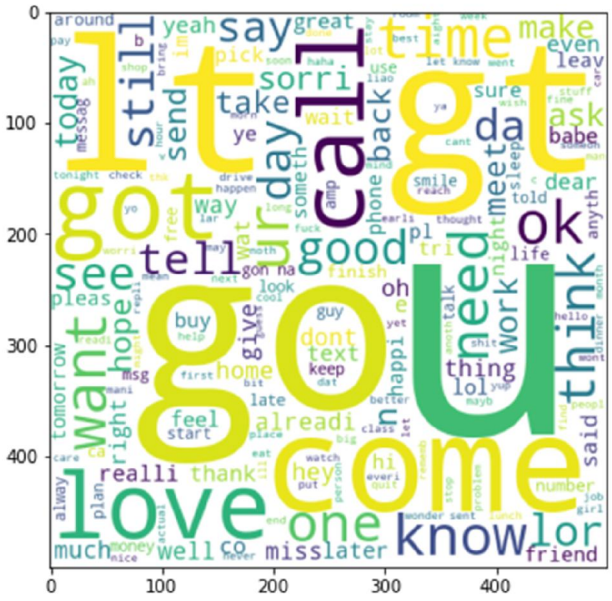

Fig 4. Ham Corpus Dataset 1

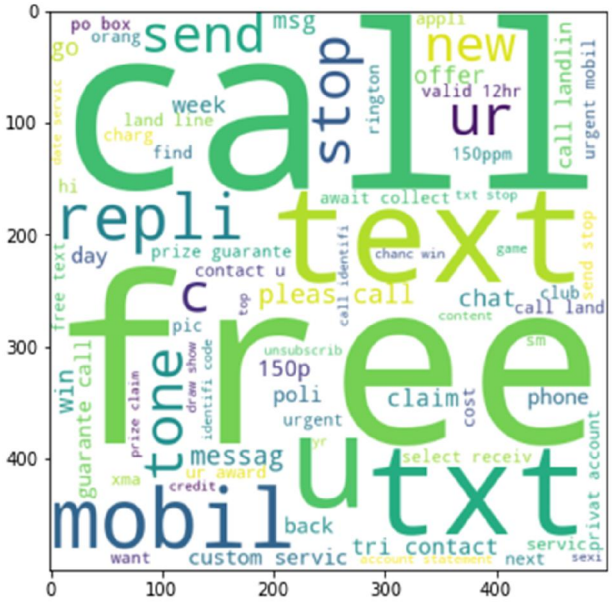

Fig 5. Spam Corpus Dataset 1

The above two WordClouds is for Dataset 1 which was downloaded from UCI Machine Learning. As you can see in ham text like ‘go’, ‘love’, ‘come’, etc are present and in spam text like ‘call’, ‘free’, ‘urgent’, etc. are present.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 5, May 2022
Impact Factor: 6.252

Fig 6. Ham Corpus Dataset 2

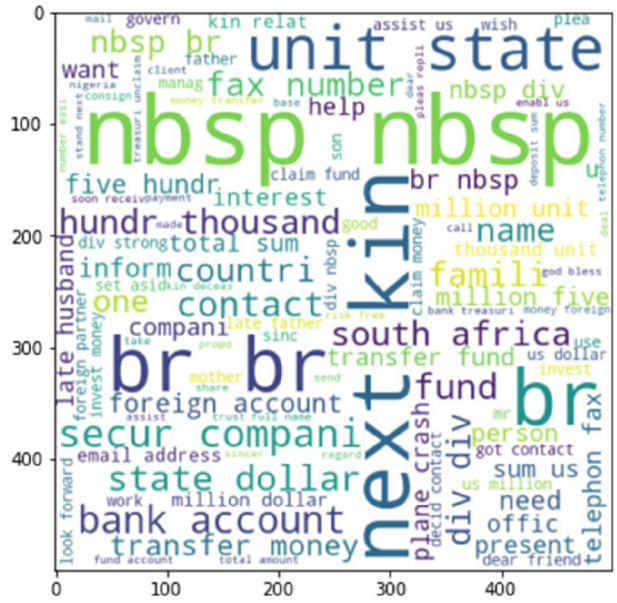

Fig 7. Spam Corpus Dataset 2

Even similar texts can be seen for dataset 2. So this WordCloud are very important for analyzing the spam texts.

IX. TEXT PRE-PROCESSING

Since machine learning techniques cannot operate directly with raw text data, sentences and words in emails must be converted into integers or vectors known as features. Feature extraction is a process of extracting a list of words from textual input and translating them into a set of features that a machine learning algorithm may use. Tokenization is accomplished in two steps. Each email is divided into two sections, with words and sentences taken separately. Words are directly entered into a database known as a data dictionary. NLTK is used to convert the sentences into vectors and store them in the data dictionary. For both corpuses, Tf-idf is a method for counting the frequency of words that appear in the dataset.The frequency was then compared to determine the outcome. For this project, we isolated the frequency of only the most commonly used 3000 words for processing.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 5, May 2022
Impact Factor: 6.252

Following this input, words are sent to a manually created function called transform text, where we change all text to lower case, remove punctuation, remove alpha-numeric values, remove stop words, and convert words into their stemmed form using a function called stemming. This step significantly reduces the system's computation size. It also aids in improved message classification.

X. ALGORITHM

Supervised Learning Method: Supervised learning methods operate on a set of data with predefined input and output. It can then forecast the outcome of the specific problem. Supervised Learning is always learning from the data it receives. Then it uses its very own probabilistic mapping system to create the outcome using the provided input.

Unsupervised Machine Learning Algorithms:

The suggested model is not pre-trained using data samples or explicit instructions, as the name indicates. As a result, there is no training for these systems. This technique's analysis is based on the dataset and reveals the group's common attributes, structures, and features. After that, the data is reorganised in a different structure or style.

Boosting Algorithm: Boosting is a strategy for merging many models to maximise the potential of this notion by attempting to discover models that complement one another. It's indeed identical to bagging in that it employs voting to categorise or average the numerical assessment to the output of a single individual model. Another similarity is because it integrates similar models, such as decision trees. On the other side, it is iterative. While bagging employs independent models, boosting employs new models that are impacted by the effectiveness of prior models. Boosting fortifies new models, allowing them to become specialists in scenarios that were previously handled improperly. Finally, the contribution of a model is evaluated by increasing its confidence rather than assigning equal weight to all models.

Support Vector Machine: Another well-known and commonly used Machine Learning classification tool is the Support Vector Machine. Some systems employed SVM as their only method of system categorization, while others used a mix of approaches, including SVM. The weights represent the relative value of several observations and analyses; 'classes.' The updated weighted SVM algorithm, according to the study, provides greater performance metrics. The SVM approach generates a hyperplane from which many classes to analyse various dataset attributes are generated. SVM is applicable to any number of vector dimensions. A two-dimensional line would be the approach. It would be a three-dimensional hyperplane.

K-Nearest Neighbor: One of the few systems that employs supervised learning to learn algorithms is K-Nearest Neighbor. The KNN set of rules implies that new case records and existing instances are comparable, and assigns the new case to the category that is closest to the existing ones. The KNN algorithm keeps all previous evidence and classifies incoming data points based on their similarity. This indicates that, regardless of how new statistics arise, they can be easily classified using the k NN method.

Naïve Bayes: This is a well-known supervised machine learning algorithm. The Bayes' rule, which aims to calculate the chance of an occurrence occurring based on even similar past knowledge and conditions, was used to construct this. This approach is very scalable, quick, and simple to implement into a system. The Naive Based algorithm considers the characteristics to be distinct. This was employed in an approach devised by to overcome the problem of random variable independence by utilising 23 distinct categorization criteria. Using a Decision Tree and Naive Based, this system obtains the desired result. This method's key limitation is that it can only be employed if the input characteristics are "totally independent of one another."

Decesion Tree: Another algorithm that has been employed more consistently in the supervised learning technique study is the decision tree machine learning algorithm. The reason for its increased use is that it is a straightforward approach with basic explanations and graphics. This approach works with both different size data sets. The system is capable of handling both numerical and category data. In their final system, they used DT in combination with other algorithms. DT was used with binomial classification of junk and normal emails in the tier three level. The model can identify spam in real time, and DT has important insights for this feature since it has a straightforward computing method, which is critical for efficient real-time computational needs.

Linear Regression: Linear regression is a simple and well-known system learning approach. It's a statistical technique for predicting outcomes. Sales, revenue, age, product costs, and other real or numeric factors are all predicted using linear

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 5, May 2022
Impact Factor: 6.252

regression. This could mean that our proposed method is more environmentally friendly than the standard classifier for all datasets.

Random Forest: Random Forest is an ensemble learning approach that use a large number of uncorrelated decision trees to produce a model that performs well on new datasets due to superior generalisation to previously unseen data. It minimises variance and always prevents over-fitting of the model by employing the Bootstrap Aggregation or Bagging strategy, in which numerous subsets of data are formed from the training set and chosen at random with replacement. Each decision tree is then trained using these subgroups. Predictions from a large number of uncorrelated models result in a strong performance on new datasets.

Extra-Trees Classifiers: This class implements a metaestimator, which uses averaging to boost projected accuracy and control over-fitting by fitting a number of randomised decision trees (also known as extra-trees) on distinct sub-samples of the dataset.

AdaBoost: AdaBoost is a computational model that regards an instance's weight as a natural number. The existence of occurrence weights is governed by how the error of a classifier is measured. It is the sum of the mistakenly categorised occurrences' weights divided by the total weight of all occurrences, not the proportion of incorrectly classified cases. When we weight occurrences, we may direct the attention of the learning algorithm to a specific group of occurences with a high weight. These incidents are critical because they must be identified correctly. The boosting method gave equal weight to every instance in the training data. To build a classifier, the learning technique reweights each instance in respect to the classifier's output.

GBDT: Gradient boosting classifiers are a set of machine learning approaches that combine several weaker models into a single, powerful model with robust and accurate output. These models are popular because they can correctly categorise datasets. Decision trees are commonly used to create models for gradient boosting classifiers. But how are the values collected, handled, and classified? The process by which a machine learning model splits data into discrete classes is known as classification. Each data collection has its own set of classes, which are categorised accordingly.

XGboost: Extreme Gradient Boosting (XgBoost) is a technique for sequentially generating decision trees. In XGBoost, weights are quite important. Weights are assigned to all of the independent factors, which are then input into the decision tree, which predicts outcomes. The results of these many classifiers/predictors are then integrated to create a more robust and accurate model. It can handle issues including regression, classification, rank, and custom prediction.

Bagging Classifier: Bagging classifiers are ensemble meta-estimators that apply basic classifiers to random subsets of the given dataset and then aggregate their individual predictions (through voting or average) to get a final prediction. A meta-estimator may sometimes be used to reduce the variance of a black-box estimator by including randomness into the building process of the black-box estimator. Each base classifier is trained in parallel using a training set formed by randomly selecting N instances (or data) with replacement from the original training dataset - where N is the size of the original training set. The training set for each base classifier is separate from the others. Some of those same data items in the testing set may be reproduced, while others are omitted.

XI. PROPOSED SYSTEM

Step 1: Import Dataset
Step 2: Analyze the Dataset
Step 3: Data Cleaning
3.1: Remove unwanted columns, duplicate rows, etc.
3.2: Rename rows for proper data processing
Step 4: Exploratory Data Analysis (EDA)
Step 5: Text Pre-Processing
- **5.1:** Remove Capital letters
- **5.2:** Remove Punctuation Marks
- **5.3:** Remove alpha-numeric values
- **5.4:** Remove Stopwords
- **5.5:** Perform Stemming

Step 6: Apply Classifiers

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 5, May 2022
Impact Factor: 6.252

Step 7: Evaluate Performance
Step 8: Pickle the best performing algorithm

XII. RESULT AND DISCUSSION

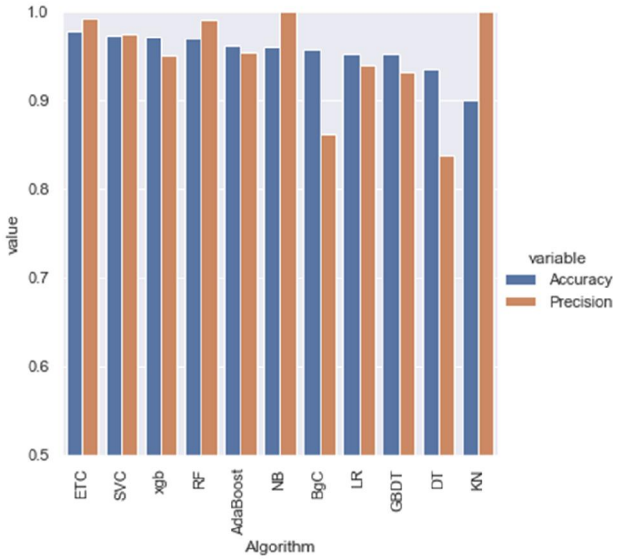

Fig 8. Performance of Classifiers (Dataset 1)

So, after executing and implementing the model and found the results for the Machine Learning algorithms, comparing all algorithms and their performances we have come to a conclusion that since for dataset 1, precision was the most important parameter because of the dataset being imbalanced so Multinomial Naïve Bayes is the best performing algorithm. Even K-NN algorithm had a precision value of 100% but its accuracy being less than Multinomial Naïve Bayes. Also Extra-Trees Classifier and Support Vector Machine have good performances but misses some bit because of precision.

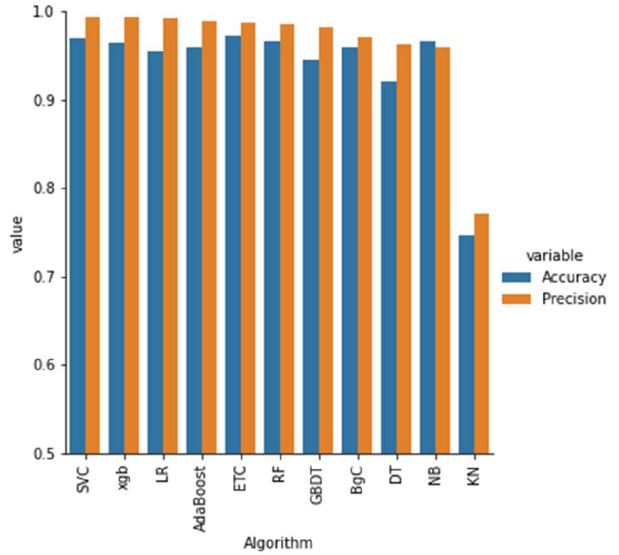

Fig 9. Performance of Classifiers (Dataset 2)

For Dataset 2, since it being more on the balanced side of things both the parameters are considered to be equally importan. So based on that Support Vector Machine, Extra-Trees Classifiers and Random Forrest were the top 3 performing Classifiers. Among these SVM has the most accuracy and precision of 96.955% and 99.439% respectively.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 5, May 2022
Impact Factor: 6.252

XIII. CONCLUSION

To apply to specific dataset we have used multiple classification techniques to tuned for their peak performance. We built an adaptive classification for Naive Bayes in the data analysis portion. You can then increase the algorithm's stability and performance in comparison to other data analysis algorithms.

REFERENCES

[1]. Chao Chen, Yu Wang, Jun Zhang, Yang Xiang, Wanlei Zhou, Statistical Features-Based Real-Time Detection of Drifted Twitter Spam, IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY, VOL. 12, NO. 4, APRIL 2017.

[2]. L. Breiman, Random forests, Mach. Learn., vol. 45, no. 1, pp. 5-32, 2001.

[3]. C. Grier, K. Thomas, V. Paxson, and M. Zhang, @spam: The underground on 140 characters or less, in Proc. 17th ACM Conf. Comput. Commun. Security, 2010, pp. 27-37.

[4]. H. Kwak, C. Lee, H. Park, and S. Moon, What is twitter, a social network or a news media? in Proc. 19th Int. Conf. World Wide Web, 2010, pp. 591-600.

[5]. K. Lee, J. Caverlee, and S. Webb, Uncovering social spammers: Social honeypots + machine learning, in Proc. 33rd Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.,2010, pp. 435-442.

[6]. J. Oliver, P. Pajares, C. Ke, C. Chen, and Y. Xiang, An in-depth analysis of abuse on twitter, Trend Micro, Irving, TX, USA, Tech. Rep., Sep. 2014.

[7]. Song, S. Lee, and J. Kim, Spam ltering in twitter using sender-receiver relationship, in Proc. 14th Int. Conf. Recent Adv. Intrusion Detection, 2011, pp. 301-317.

[8]. K. Thomas, C. Grier, D. Song, and V. Paxson, Suspended accounts in retrospect: An analysis of twitter spam, in Proc. ACM SIGCOMM Conf. Internet Meas. Cof., 2011, pp. 243-258.

[9]. C. Yang, R. Harkreader, and G. Gu,Empirical evaluation and new design for fighting evolving twitter spammers, IEEE Trans. Inf. Forensics Security, vol. 8, no. 8, pp. 1280-1293, Aug. 2013.

[10]. S. Yardi, D. Romero, G. Schoenebeck, and D. Boyd, Detecting spam in a twitter network, First Monday, vol. 15, nos. 1-4, pp. 1-13, Jan. 2010.