wjert, 2022, Vol. 8, Issue 6, 175-193. *Research Article* ISSN 2454-695X

World Journal of Engineering Research and Technology

WJERT

www.wjert.org SJIF Impact Factor: 5.924

ANALYSIS OF SPAM DETECTION MODELS USING MACHINE LEARNING

Prakash Mani Badal*

Dept. of Computer Science and Engineering, NIT Jamshedpur.

Article Received on 22/04/2022 Article Revised on 12/05/2022 Article Accepted on 01/06/2022

***Corresponding Author**
Prakash Mani Badal
Dept. of Computer Science and Engineering, NIT Jamshedpur.

ABSTRACT

Messages have become an important component of our daily communication in today’s society, and the number of messages received by a user has increased significantly. There has been a large increase in the amount of spam messages in tandem with the rise in message volume. Spam communications are unsolicited mass messages that are sent to a large number of people. As a result, we get a lot of spam. These spam messages frequently obscure essential messages and may contain dangerous links that lead to harmful websites, jeopardising users’ security. The bulk of spam classifiers are created using a single machine learning method, which may not be sufficient for correctly identifying a message as spam or ham. This thesis adds to the development of a Hybrid Spam classifier through the use of ensemble machine learning. In our research, a total of seven machine learning algorithms were shortlisted. These were then utilised to create 7 basic classifiers using a single method, 35 level 3 hybrid models using three algorithms, 21 level 5 hybrid models using five algorithms, and one level 7 hybrid model using seven algorithms. To narrow down the best hybrid models, a total of 64 models were trained and assessed.

CHAPTER 1

INTRODUCTION

1.1 Spam

Spam is defined as unsolicited bulk messages that are of no utility to the recipient. Spamming[1] is the practise of sending such communications in bulk to numerous domains such as emails, messages, social media platforms, or any other form of com- munication and on the Internet.It’s an act of exploiting the electronic communica- tions system. In the twenty-first century, we live in an electronically connected world. Emails and text messages have surpassed all other modes of communication as the primary means of communication. However, it is really regrettable that these ex- tensively used communication methods have been routinely misused to compromise information confidentiality, integrity, and availability.Spamming via email, commonly known as spam, has become one of the most common methods of spamming in recent generations. Spamming not only reduces the availability of information by obscuring critical messages, but it also directs users to harmful websites, putting their security at risk. A spam email can be categorised as[2] according to the general definition of spam.

1. A message that the user has not requested.
2. A piece of mail with commercial worth.
3. An e-mail that has been sent in bulk.

These emails are undesired, and they take up a lot of bandwidth and storage space for consumers. It also wastes users’ time because they must open and close it. The goal of this study is to offer a spam detection technique that will correctly identify a communication as spam or ham.

1.2 History of Spam

The origins of spam messages may be traced back to the mid-1990s, when the Internet was initially used for commercial purposes and people, particularly marketers and publicists, experimented with it to see what was feasible. It didn’t take long for people to realise that Email Spam has become an inevitable and repeated process.

Several hundred users of the Arpanet, the world’s first internet, got the first email spam on May 3, 1978, which contained an advertisement for Digital Equipment Cor- poration.[2]

In April 1993, the term”spam” was coined for the first time. People began making fun of the situation after a user unintentionally sent 200 emails to a news agency. This was the first time a person used the term spam for a batch of messages instead of Email. On January 18, 1984, there was the first large-scale spam attack. when. Every single newsgroup received a mail with the subject ”Jesus Is Coming.”

1.3 Effects of Spam

These unwanted bulk messages are not only annoying because they flood users' inboxes and send them regular reminders of incoming texts, but they can also be dangerous. The source of these communications is usually anonymous, and spam messages have a number of harmful repercussions.

1. Direct Consequences: Unfortunately, spam has become a route for selling counterfeit goods and putting harmful software on customers' PCs. As a result, a variety of frauds have occurred, and you've been a victim of a breach in cybersecurity.
2. Wastage of Resources: Spam communications are transmitted in bulk, causing traffic, and because they are unwanted, they waste valuable bandwidth and storage space.
3. Wastage of Human resource: Not only does it waste resources, but it also wastes human labour. Users are finding it difficult to find useful messages due to the large amount of undesirable bulk emails.

Spam wastes a lot of storage space by filling users' inboxes with unsolicited mes- sages, and it also has a negative impact on the network band, resulting in a bad network experience.

1.4 Anti Spam Measures

Although no anti-spam approach is 100% effective in preventing spam messages, the amount of spam messages can be greatly decreased by using the tactics listed below.

i. End user techniques: To limit the quantity of spam communications received, end users should remember the following procedures.

(a) Discretion: Users should not share their email address with strangers or un- known groups. They should only give out their email address to people they know and trust.

(b) Address Munging:Address munging is a method of masking a user's email address, such as ”no-one@example.com” into ”no-one at example dot com.”

(c) Report Spam Messages:End users should report spam messages so that administrators can mark them as spam. It also aids in the creation of more accurate spam data sets.

(b)Disposable Email address:Though it is rarely recommended, providing an email address to third-party websites is frequently needed. In these circumstances, a different spam detection method should be utilised.

2. Automated technique for email administrator:Administrators can use the measures listed below to reduce spam communications.

(a)Authentication: Spammers aim to impersonate a credible source by spoofing their address and ties. SPF, DKIM, and DMARC are examples of systems that make spoofing more difficult.

(b)Challenge Response system: Before accepting an unknown sender, the administrator can set up a series of obstacles to ensure that the sender is genuine. This strategy reduces the impact of Spambot, a widely used spam distribution tool.

(c)Checksum Based Filters: Spam communications are typically sent in large batches with slight differences. As a result, spam communications are more likely to share a Checksum. This vulnerability is used by a checksum-based filter to detect spam messages.

(d)DNS-Based Blacklisting:A DNS-based blacklist can be used by an admin- istrator to reject a message whose address matches the blacklist.

(e)URL based Filtering:Most spam communications contain a URL, which is collected and checked against the Domain Block List in this technique.

(d)Rule Based Filtering: Here, the emails are scanned for a list of words and regular expressions that have previously been detected in spam emails.

3. New Solution and Ongoing Research

(a)Cost Based System : The idea here is to charge a fee for sending emails. Because spam messages are delivered in bulk, spammers will be unable to transmit spam messages due to the high expense of doing so.

(b)Machine Learning Based approach: Binary classification is possible with a number of machine learning algorithms. This aspect of the Machine learning-based technique can be used to determine whether a communication is spam or ham.

1.5 Machine learning Based Spam Detection

Spam messages and spam emails have increased exponentially with the emergence of mobile phones and the Internet. Spam communications are unsolicited bulk messages that pose a serious threat to the CIA's (Confidentiality, Integrity, and Availability) triad . As a result, spam messages flood inboxes, and vital messages known as ham are concealed, resulting in the denial of service to legitimate users. Spam communications frequently contain hazardous links that direct clients to malicious websites, leaving themexposed to attack.[3] Both emails and text messages have become popular modes of communication, and they are widely used by a large number of people. They provide attackers with a platform to propagate spam messages, which can lead to various sorts of cyber-attacks.[4] It's critical to classify spam and ham messages as spam and ham messages to avoid cyber-attacks caused by spam mailings

(Non-Spam).There are two methods for categorising the messages. We can classify the messages using knowledge engineering or machine learning. The classification criteria in knowledge engineering must be updated on a regular basis, which is a significant difficulty.

Machine Learning has exploded in prominence in recent decades for message clas- sification. The availability of a large number of data sets, as well as the flexibility of machine learning, are the primary reasons for its growing popularity. The findings can be estimated more correctly with the help of machine learning[5] without having to programme them explicitly. The Machine learning-based classifier's steps are outlined below.

1. Choosing a Data set: Supervised Learning Algorithms are used by the vast majority of machine learning algorithms that perform binary classifications. It's critical to train our data sets when using supervised learning algorithms.
2. Pre-Processing of Data: The data sets available are in human-readable format and must be pre-owned in order to be used for model training.

The popular steps required for pre-processing are.

(a)Stop word removal

(b)Stemming

(c)Tokenization

(d)IgMatrix

Model Selection: The heart of any Machine learning-based classifier is the model used in it. The user must select the correct machine learning algorithm to classify a message as spam or ham.

Testing and Training: The models are trained with a part of the data set and the later part is used for testing. The division of Data-set is open for users. In our, work Data-set was split into a 50-50 ratio. half of the data was used for training and another half of the data was used for testing.

1.6 Motivation

The goal of this work is to look at the existing Machine Learning techniques that are extensively used to classify messages as Ham or Spam, and then use those methods to create a Hybrid solution for correctly identifying messages as Ham or Spam. Standard spam detection techniques aren't always accurate in determining whether a message is spam or ham. Traditional classifiers use a single method at a time and may be effective at categorising a certain type of message, such as spam or ham. As a result, there is much space for development in this area. the spam message identification process.

1.7 Overview of the Propose work

Based on the foregoing facts, it was evident that a single algorithm wasn't adequate to accurately categorise a message as spam or ham, and that a Hybrid classifier based on the principle of Ensemble learning was required to do so. To address the problem, we first identified seven techniques that were used to classify the message as spam or ham, and then used them to create Hybrid classifiers. We developed all of the feasible combinations of level 3 (using three algorithms), level 5 (using five algorithms), and level 7 (using seven algorithms) Hybrid classifiers in our research. There were 64 classifiers constructed in total: 7 base classifiers, 35 level 3 classifiers, 21 level 5 classifiers, and 1 level 7 classifier.

1.8 Objectives of the Thesis

1. To study and analyze the existing Machine learning algorithms and shortlist the most popular algorithms in use.
2. With help of the existing algorithms built all the possible Hybrid models.
3. Provide an in-depth comparison of all the models and propose the best model.

CHAPTER 2

Machine learning Algorithm

2.1 What is Machine Learning

Machine Learning (ML) is a subcategory of artificial intelligence that describes the process through which computers gain pattern recognition, or the ability to continu- ously learn from and make predictions based on data, and then make improvements without being explicitly programmed to do so.

2.2 How does Machine Learning work?

Machine learning is extremely complicated, and how it functions changes based on the goal and the technique employed to complete it. A machine learning model, on the other hand, is a computer that examines data for patterns and then uses those insights to better execute its assigned task. Machine learning can automate any operation that requires a set of data points or rules, including more complicated tasks like answering customer service calls and analysing resumes.

Machine learning algorithms use varying degrees of human assistance and reinforce. ment depending on the situation. Unsupervised learning, semi-supervised learning, and reinforcement learning are the four major machine learning models.

In supervised learning, the computer is given a labelled set of data to help it learn how to perform a human skill. This is the simplest model because it tries to mimic human learning.

Unsupervised learning occurs when a computer is given unlabeled data and is asked to extract previously unknown patterns or insights. Machine learning algorithms accomplish this in a variety of ways, including:

Clustering is the process of a computer identifying comparable data points in a data set and grouping them together (forming"clusters").

Density estimation is a technique in which a computer deduces information from the distribution of a data collection.

Anomaly detection is when a computer detects data points that are considerably different from the rest of the data in a data set.

PCA(Principal component analysis) is a technique in which a computer analyses and summarises a data set such that it can be utilised to create accurate predictions. In semi-supervised learning, the computer is given a set of partially labelled data and is asked to execute a task using the labelled data to figure out how to interpret the unlabeled data.

With reinforcement learning, the computer monitors its surroundings and uses the information to determine the best behaviour to reduce risk and increase reward. This is an iterative method that necessitates the use of a reinforcement signal to assist the computer in determining the best course of action.

2.3 Machine Learning Classifiers

What is classification?

The technique of predicting the class of given data points is known as classification. Targets, labels, and categories are all terms used to describe classes. The task of approximating a mapping function (f) from discrete input variables (X) to classification predictive modelling (y).

Spam detection in email service providers, for example, can be classified as a classi- fication issue. Because there are only two types of spam, this is a binary classification. To understand how given input variables relate to the class, a classifier uses some training data. In this situation, the training data must consist of known spam and non-spam emails. When the classifier has been properly trained, it can be used to identify unfamiliar emails.

Classification is a type of supervised learning in which the input data is also deliv- ered to the objectives. Classification has numerous uses in a variety of fields, including credit clearance, medical diagnosis, and target marketing.

There are two types of learners in classification as lazy learners and eager learners. 1.Lazy learners Lazy students simply save the training data and wait for the testing data. When this happens, classification is performed using the most closely related data from the stored training data. Lazy learners spend less time training but more time forecasting than motivated learners.
Ex. k-nearest neighbor, Case-based reasoning

2. Eager learners Before receiving data for classification, eager learners develop a classification model based on the available training data. It must be capable of committing to a single hypothesis that encompasses the entire instance space. Because of the model's design, eager learners require a long time to train and even longer to predict.
Ex. Decision Tree, Naive Bayes, Artificial Neural Networks

Classification algorithms

There are numerous categorization algorithms available today, and it is impossible to determine one is superior to the others. It is dependent on the application and the nature of the data collection supplied. If the classes are linearly separable, linear classifiers such as Logistic regression and Fisher's linear discriminant can outperform complicated models.

CHAPTER 3

LITERATURE SURVEY

Here we summarize related works in the field of spam detection using machine learning related to our literature.

Asif Karim; Sami Azam; Bharanidharan Shanmugam;Krishnan Kannoorpatti[6] Spam detection, spam email, and spam filtering have all been employed. The overall purpose of the study is to create an unsupervised framework that relies only on unsupervised techniques, employing a clustering strategy that incorporates numerous algorithms and largely relies on the email content (body) and subject header. Clustering was performed on a new binary dataset of 22,000 entries of ham and spam emails, which included ten features (reduced from eleven to ten after the feature reduction). Seven of the 10 features are new to this study, and were designed to reflect important analytical email characteristics from many angles. OPTICS produced the best clustering out of five different clustering algorithms studied in this study, with an average efficacy of 0.26 percent higher.

Yafeng Ren; Donghong Ji[7] have used S Deceptive opinion spam, deceptive review, machine learning, feature engineering, natural language processing, deep learning.

Aaisha Makkar; Sahil Garg; Neeraj Kumar; M. Shamim Hossain; Ahmed Ghoneim; Mubarak Alrashoud[8] Spam detection in IoT is proposed using a Machine Learning framework. Five machine learning models are assessed using multiple metrics and a vast collection of input feature sets in this framework. Each model uses the refined input attributes to calculate a spam score. This score represents the trustworthiness of an IoT device based on a variety of factors. The proposed technique is validated using the REFIT Smart Home dataset. The acquired findings demonstrate that the proposed method is more effective than other current schemes.

Jaeun Choi,Chunmi Jeon[9] Expert decision making, machine learning, real-time spam detection, social networking, and Twitter spam have all been deployed. This research proposed a complex method for detecting spam tweets that combines ex- pertise and machine learning algorithms. The effort on the expert can be reduced because many regular tweets can be filtered out in the first stage using a cost-based machine learning filter. The accuracy can potentially be increased by having profes- sionals analyse the questionable tweets. The experimental results suggest that the proposed strategy outperforms traditional machine learning techniques in terms of spam detection.

HAYOUNG OH[10] In the comment data from popular music videos - Psy, Katy Perry, LMFAO, Eminem, and Shakira - we used Decision Tree, Logistic regression, Bernoulli Nave Bayes, Random Forest, Support vector machine with linear kernel, Support vector machine with Gaussian kernel) and two ensemble models (Ensemble with hard voting, Ensemble with soft voting) combining these techniques. He splits the data into 70 percent training and 30 percent test. Then, ten machine learning al- gorithms are deployed, with five assessment measures: Acc (Accuracy rate), SC (Spam caught rate), BH (Blocked ham rate), F1-score, and MCC (Matthews correlation coef- ficient). Table 4 provides the foundation for all formulas. As a consequence, the ESM-S model performed best in Acc, SC, F1-score, and MCC, while the ESM-S model per- formed second. The ESM-S model outperformed the ANN model in SC and the NB-B model in BH in terms of accuracy, F1-score, and MCC in both datasets. Furthermore, the data set containing 1,000 spam and 1,000 legitimate comments outperformed the control group.

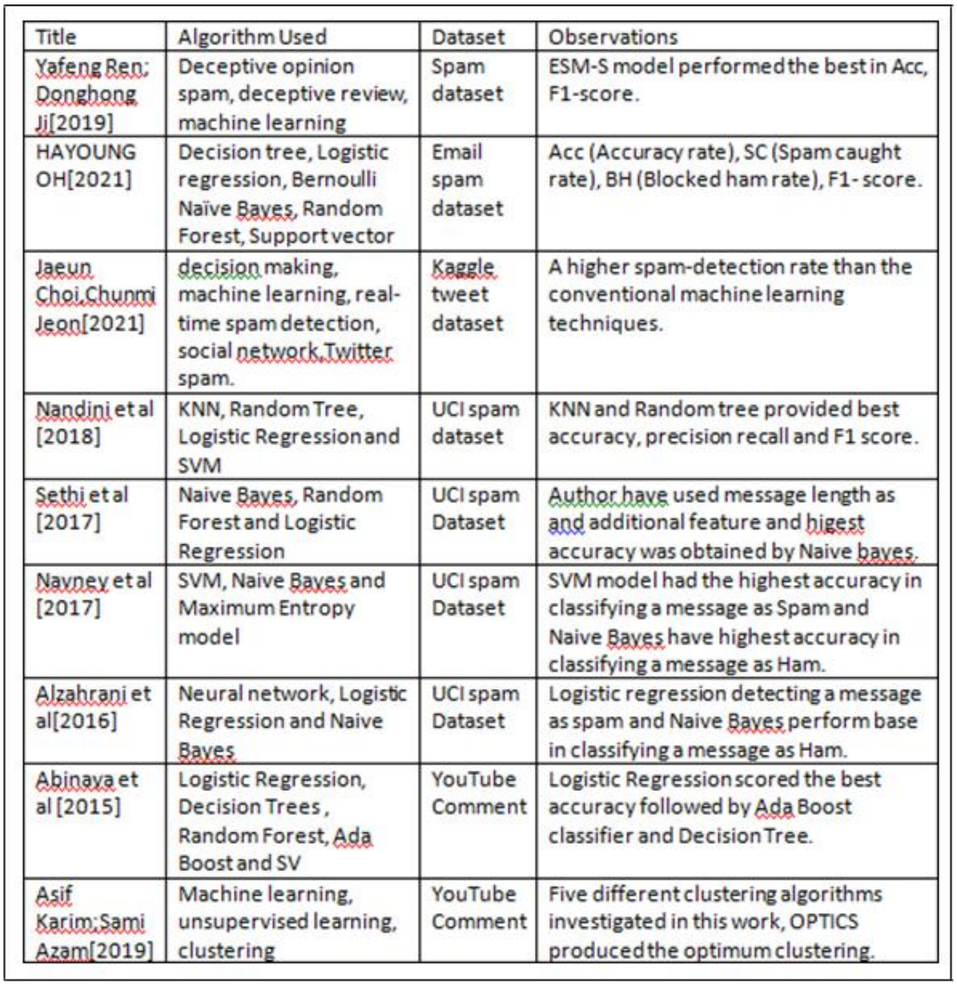
Nandini et al [?] have used KNN, Random Tree, Logistic Regression and SVM Classifier for Email Spam Detection. The author used the UCI spam dataset counting 5774 messages to train and test the data. The author used accuracy, Precision, Recall and F1 score to evaluate a message was spam or not. When it came to accuracy KNN and Random tree scored the highest accuracy of 0.999348 followed by Logistic Regression with an accuracy of 0.931319. SVM scored an accuracy of 0.907629 and the Naive Bayes classifier produced an accuracy of 0.795262. When it came to Precision Random Tree and KNN scored the highest Precision of 0.99999 followed by Logistic Regression with a Precision of 0.931. Naive Bayes Scored 0.845 in Precision and SVM scored 0.117 which was the Least of all the classifiers. The next metrics that were used for comparison was recall in which again Random Forest and KNN scored the highest with 0.9999 recall followed by Logistic with a recall of 0.931. SVM scored a recall of 0.908 and Naive Bayes scored a recall of 0.795. IN F1 score again KNN and Random tree scored 0.999, Logistic Regression scored 0.931, SVM scored 0.907 and 10 Naive Bayes scored F1 score of 0.797. Sethi et al have used Naive Bayes, Random Forest and Logistic Regression for detecting spam using various Machine learning algorithms. The author has used the UCI spam data set to train and test the models to classify a message as spam or ham. The author used the feature of message length in the training of the dataset. The author observed the average length of spam message was 176 characters and the average length of Ham message was 56 characters. The author did a comparative study in finding the accuracy of the models between using the message length and without using the message length. It was found that using the message length as a feature in classifying a message as Spam or Ham performed better than without using the message length as spam or ham. In which Na¨ıve Bayes scored 98.445 accuracy random forest scored 97.54 accuracy and logistic regression scored 95.454. Navney et al [?] have used SVM, Naive Bayes and Maximum Entropy model.

The author has used The UCI Spam dataset for training and testing of data as Spam or Ham. The author found SVM Classifier had an accuracy of 0.984 in successfully predicted a message as Ham had an accuracy of 0.964 in predicting a message as spam leading to an overall accuracy of 0.974. It was followed by the Naive Bayes classifier which had an accuracy of 0.909 in successfully classifying a message as spam and an accuracy of 0.982 in successfully classifying a message as Ham and scored an overall accuracy of 0.9455. Max Entropy classifier scored an accuracy of 0.98 in classifying a message as Ham and 0.859 in classifying a message as spam the overall accuracy of the model was found to be 0.9195. Alzahrani et al [?] have done a comparative study of Machine learning algorithms and have used Neural network, Logistic Regression and Naive Bayes to classify a message as spam or ham. The author has also used the UCI spam dataset for the 11 training and testing of data. It was found that Neural network algorithms presented the best accuracy, which was nearly 98.0 it also had the least error in detecting a message as spam as it wrongly classified only two messages out of the total of 136 spam messages it was followed by logistic regression with an accuracy of 94.26. logistic Regression also had a higher error rate in classifying a message as Ham as Out of 1019 ham messages around 63 messages were wrongly classified. However, when it came to the error rate for classifying a message as spam it had a very less error rate as only 1 of the 96 spam messages were wrongly classified.

Na¨ıve Bayes scored an accuracy of 88.16. but it also had the lowest error rate in classifying a message as Ham as out of a total of 881 Ham messages only 28 messages were wrongly classified. However, it had a higher error rate in classifying a message as spam as out of a total of 234 spam messages it wrongly classified 104 messages which were relatively very high. Abinaya et al [?] have done an analysis of spam detection on social media platform. The author has taken youTube comment as the dataset and has performed spam detection on Youtube comment detection. The author has used Logistic Regression, Decision Trees Classifier, Random Forest, Ada Boost Classifier and Support Vector Machine for Classifying spam on social media platforms. In which Logistic Regression scored an accuracy of 0.9540 followed by Ada Boost Classifier and Decision Tree Classifier with an accuracy of 0.9438. Support Vector machine-scored an accuracy of 0.5051 and Random Forest achieved an accuracy of 0.8469.

3.1 OBSERVATION

The key observations of the Literature survey are listed in table



Title	Algorithm Used	Dataset	Observations
Yafeng Ren; Donghong Ji[2019]	Deceptive opinion spam, deceptive review, machine learning	Spam dataset	ESM-S model performed the best in Acc, F1-score.
HAYOUNG OH[2021]	Decision tree, Logistic regression, Bernoulli Naïve Bayes, Random Forest, Support vector	Email spam dataset	Acc (Accuracy rate), SC (Spam caught rate), BH (Blocked ham rate), F1- score.
Jaeun Choi,Chunmi Jeon[2021]	decision making, machine learning, real-time spam detection, social network,Twitter spam.	Kaggle tweet dataset	A higher spam-detection rate than the conventional machine learning techniques.
Nandini et al [2018]	KNN, Random Tree, Logistic Regression and SVM	UCI spam dataset	KNN and Random tree provided best accuracy, precision recall and F1 score.
Sethi et al [2017]	Naive Bayes, Random Forest and Logistic Regression	UCI spam Dataset	Author have used message length as and additional feature and higest accuracy was obtained by Naive bayes.
Navney et al [2017]	SVM, Naive Bayes and Maximum Entropy model	UCI spam Dataset	SVM model had the highest accuracy in classifying a message as Spam and Naive Bayes have highest accuracy in classifying a message as Ham.
Alzahrani et al[2016]	Neural network, Logistic Regression and Naive Bayes	UCI spam Dataset	Logistic regression detecting a message as spam and Naive Bayes perform base in classifying a message as Ham.
Abinaya et al [2015]	Logistic Regression, Decision Trees , Random Forest, Ada Boost and SV	YouTube Comment	Logistic Regression scored the best accuracy followed by Ada Boost classifier and Decision Tree.
Asif Karim;Sami Azam[2019]	Machine learning, unsupervised learning, clustering	YouTube Comment	Five different clustering algorithms investigated in this work, OPTICS produced the optimum clustering.

Figure 3.1: Observation of various models.

CHAPTER 4

PROPOSED WORK

In this section, we broadly described the solution proposed by us. A Flow diagram of the proposed work is shown in figure.

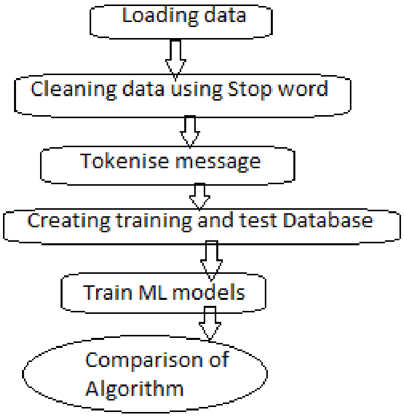


Figure 4.1: Flow chart of proposed work.

4.1 Need for Pre-Processing of Text

The data in the datasets which are used in building the classifiers are labelled datasets a spam message is labelled as 1 whereas a ham message is labelled as 0.A few of theexample consisting of Ham messages are shown in table 4.1 and spam messages shown in table 4.2.

Table 4.1: Ham Messages.

Subject: enron methanol; meter
Subject: hpl nom for january 9, 200 see
Okay I can try, but cannot commit.
I am good too. Yes weekdays are busy, all thanks to office.

Table 4.2: Spam Messages.

Subject: photoshop, windows, office. cheap
Post Diwali offer! Get 30percent off
Get a whole-some Chocolate Shake free on orders above Rs. 2000.

When we look at the Spam and Ham messages closely, we can observe that they are significantly different. The Ham messages are more personal, as if they were sent to a specific individual. Spam messages, on the other hand, appear to be quite general and were intended for widespread distribution. Aside from that, many phrases like offer, sale, and free are frequently utilised in spam communications. To classify a message as spam or not, we'll look for these words in spam messages. Computers, on the other hand, are unable to read and analyse text in the same way that humans can.We must pre-process the texts in a way that our computers can understand in order to draw relevant information from the data sets. The transforming of text into a form that a computer can interpret is known as text preprocessing. Pre-processing our texts will be done with the Natural Language Tool Kit. NLTK is a popular programming language for working with human language data, and it's written in Python. Natural languages are transformed into something that computers can understand with the help of NLTK.

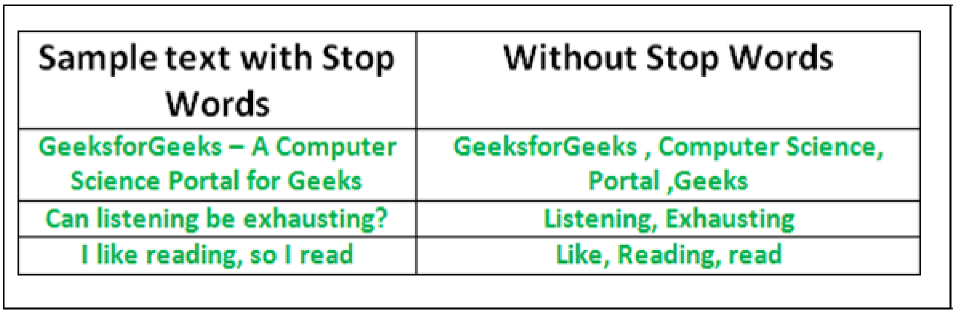
4.2 Removal of Stop word

The process of converting data to something a computer can understand is referred to as pre-processing. One of the major forms of pre-processing is to filter out useless data. In natural language processing, useless words (data), are referred to as stop words.

About Stop words?

A stop word is a widely used word (such as ”the,” ”a,” ”an,” or ”in”) that a search engine has been configured to disregard while indexing and retrieving entries as the result of a search query. We don’t want these terms to eat up important storage space or processing time in our database. We may easily remove them by keeping a list of terms that you regard to be stop words. In Python, the NLTK (Natural Language Toolkit) has a list of stopwords in 16 different languages. They are located in the nltk data directory.

To check the list of stopwords you can type the following commands in the python shell.
import nltk from nltk.corpus import stopwords



Sample text with Stop Words	Without Stop Words
GeeksforGeeks – A Computer Science Portal for Geeks	GeeksforGeeks , Computer Science, Portal ,Geeks
Can listening be exhausting?	Listening, Exhausting
I like reading, so I read	Like, Reading, read

Figure 4.2: Table print(stopwords.words(’english’))

4.3 What is Stemming

The process of developing morphological variants of a root/base word is known as stemming. Stemming algorithms or stemmers are terms used to describe stemming programmes. The phrases ”chocolates,” ”chocolatey,” and ”choco” are reduced to the root word ”chocolate,” and ”retrieval,” ”retrieved,” and ”retrieves” are reduced to the stem ”retrieve.” In natural language processing, stemming is an integral aspect of the pipelining process. Tokenized words are fed into the stemmer. What is the source of these tokenized words? Tokenization is the process of breaking down a document into individual words. To learn more about tokenization and how it works, see this article:

Some more example of stemming for root word ”like” include: ”likes”
”liked”
”likely”
”liking”

Errors in Stemming.

There are mainly two errors in stemming –
over-stemming:

When two words are stemmed from the same root but have different stems, this is known as over-stemming. False-positives are another term for over-stemming.
under-stemming:

When two words with the same root but distinct stems are stemmed together,this is known as under-stemming. False-negatives are the result of under-stemming.
Applications of stemming :

Information retrieval systems, such as search engines, use stemming. It is used in domain analysis to determine domain vocabularies.

4.4 Tokenisation

Tokenization is the process of dividing something into little tokens. We can easily tokenize our datasets with the help of NLTK. Sentence tokenizer and word tokenizer are two of the most used tokenizers. We tokenize our document in the form of sentences in sentence Tokenizer. For example, if we have a document with 5 sentences, our sentences will be tokenized in 5 sections and we may utilise them separately. The elements of our array will be each sentence. Our complete dataset will be separated into words instead of sentences, and each word will be a member of the array. After stopping word removal and stemming, we will be left with a Bag of words containing unique words in their base form.

4.5 IG Matrix

Because we need to identify the words in a sentence for classification, we encode our data set as a matrix, with each row containing a Bag of words. In the rows, all of the unique terms are displayed, and in the column, all of the spam messages are represented. For example, suppose we have 10 spam messages and after stopping words and stemming, the number of unique words is 100. In the IG matrix, to be precise.

CHAPTER 5

SIMULATION AND RESULTS

We have used Python for the training of Data sets. NLTK was used for pre-processing of data sets. We have used sci-kit learn was used for the machine learning library.

5.1 Data Set used

We have used the spam ham dataset set contributions. The dataset was downloaded from the kaggle learning repository. It consists of a total of 20684 messages. The messages in the dataset are labelled as 1 and 0 with 1 depicting a message as spam and 0 depicting the message as ham. Each line consists of one message and is composed of two columns with the first column holding a numeric value 1 or 0 depending upon the context of the message and the second column consist of Raw data. The dataset was split into a 75:25 ratio with 75 percent of the dataset for training purpose and the remaining was used for testing purpose.

5.2 Results and comparisons

In this section, we compare the results obtained from Base classifiers, Hybrid Classifier of Level 3, Hybrid Classifier of level 5 and Hybrid Classifier of Level 7. For comparison, we have compared the classifiers giving the best performance for a particular Metrics

1. Accuracy Fig shows the comparison of highest accuracy obtained in different level.

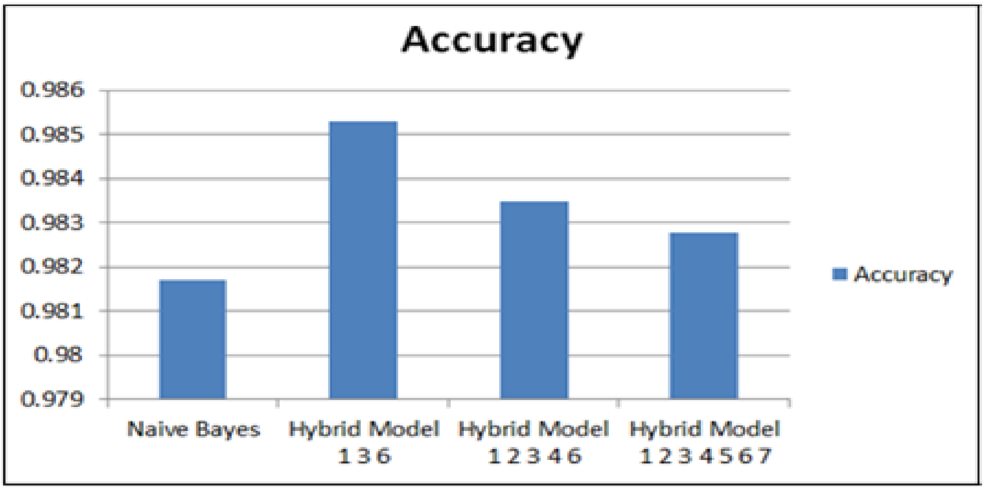


Figure 5.1: Accuracy of various models.

Figure 5.1 shows a comparison of Accuracy between various models. We can clearly see the proposed hybrid models have significantly out performed the traditional clas- sifiers with highest accuracy of 0.985284 which was obtained by Hybrid Model 1 3 6 built using Navies Bayes Classifier, Support Vector Machine and Decision Tree as it’s underlying base models.

Figure 5.2 shows a comparison of Precision between various models.Here also We can clearly see the proposed hybrid models have performed better than the traditional classifiers. The highest Precision of 0.991488 was achieved by Hybrid Model 12 4 built using built using Navies Bayes Classifier, Logistic Regression and Random Forest as it’s underlying Base models.

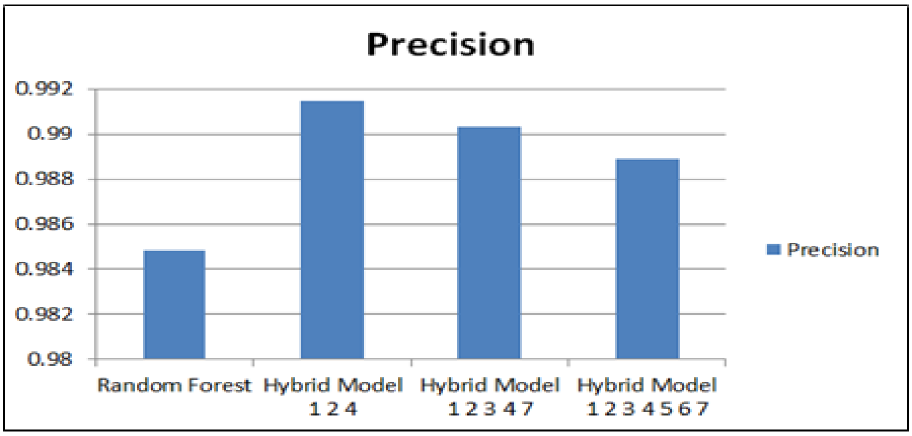


Figure 5.2: Precision of various models.

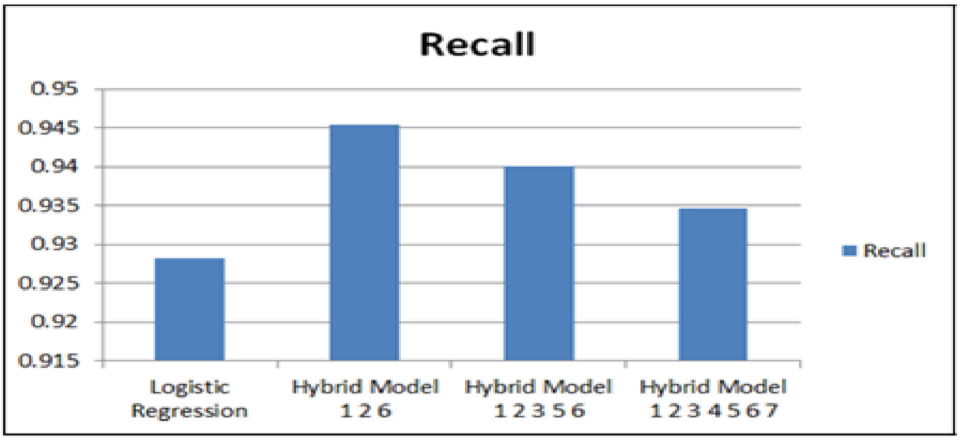


Figure 5.3: Precision of various models.

A comparison of Recall between various models is shown in Figure 5.3. We can see that in general, the proposed hybrid models outperformed traditional classifiers. Random Forest and Hybrid Model 1 2 6 constructed with Navies Bayes Classifier, Logistic Regression, and Decision Tree had the greatest Recall of 0.945365.

CHAPTER 6

CONCLUSION AND FUTURE WORK

We investigated spam detection using Machine Learning in this research. We tried to narrow down the best classifiers for determining if a message is Spam or Ham. We then combined the classifiers to create Hybrid Models. We ended up attempting all of the options.

Combinations of Level 3, Level 5, and Level 7 Hybrid Classifiers are possible.

In terms of performance, three classifiers developed using three classifiers as foun- dation models fared the best.

Accuracy, Precision Score are all important factors to consider. As a result, Hybrid can be concluded.

Our goal for future work is to contribute to greater improvements in spam detection. When it comes to supervised learning, the most significant disadvantage is that it requires prior training. To counter this disadvantage, we would like to investigate the possibilities of unsupervised learning and the many groupings that it enables. We're also excited to use Neural Networks to identify messages as Ham or Spam, and we're open to developing a hybrid model that uses both supervised and unsupervised learning.

ACKNOWLEDGEMENT

It brings me great pleasure to convey my heartfelt appreciation to Prof. Danish Ali Khan, my supervisor, for his invaluable direction, motivation, and ongoing inspiration, as well as for their ever-cooperative attitude, which has enabled me to complete my thesis in its current form.

My heartfelt gratitude also goes to Dr. Sanjay Kumar, Head of Department of Computer Science and Engineering for providing me the opportunity to avail the ex- cellent facilities and infrastructure. I am equally thankful to Prof. D K Yadav, Dr. Ashok Kumar Mehta, Dr. C. Azad, Dr. Dilip Kumar Shaw, Dr. Alekha Mishra and all non-teaching staff of Department of Computer Science and Engineering for their guidance and support.

I am also thankful to my batch-mates who encouraged me to achieve this target. I am also thankful to all my family members whose love, affection, blessings and patience encouraged me to carry out this thesis successfully. I also extend my gratitude to all my friends for their cooperation.

Finally, once again, I thank Almighty God, my lord for giving me the will power and strength to make it happen.

BIBLIOGRAPHY

1. Jaeun Choi and Chunmi Jeon. Cost-based heterogeneous learning framework for real-time spam detection in social networks with expert decisions. IEEE Access, 2021; 9: 103573–103587.
2. A Ali. Visualizing the social media universe in 2020. Retrieved on May, 16: 2020, 2020.
3. Elisa Bertino and Nayeem Islam. Botnets and internet of things security. Com- puter,

2017; 50(2): 76–79.

4. Ali Dorri, Salil S Kanhere, Raja Jurdak, and Praveen Gauravaram. Blockchain for iot security and privacy: The case study of a smart home. In 2017 IEEE international conference on pervasive computing and communications workshops (PerCom workshops), pages 618–623. IEEE, 2017.
5. Lidia Dobrescu, Serban Obreja, Marius-Constantin Vochin, Dragos, Dobrescu, and Stela Halichidis. New approaches for quantifying internet activity. In 2019 E- Health and Bioengineering Conference (EHB), pages 1–4. IEEE, 2019.
6. Onur Varol, Emilio Ferrara, Clayton Davis, Filippo Menczer, and Alessandro Flammini. Online human-bot interactions: Detection, estimation, and character- ization. In Proceedings of the international AAAI conference on web and social media, volume 11, 2017.
7. Soroush Vosoughi, Deb Roy, and Sinan Aral. The spread of true and false news online. Science, 2018' 359(6380): 1146–1151.
8. Niddal H Imam and Vassilios G Vassilakis. A survey of attacks against twitter spam detectors in an adversarial environment. Robotics, 2019; 8(3): 50.
9. R Lerman and H Denham. charged in massive twitter hack, including alleged teenage 'mastermind'. Washington Post. Retrieved, 2020; July 31: 3.
10. Clayton Allen Davis, Onur Varol, Emilio Ferrara, Alessandro Flammini, and Fil- ippo Menczer. Botornot: A system to evaluate social bots. In Proceedings of the 25th international conference companion on world wide web, 2016; 273–274.