# DIMENSION REDUCTION FOR SPAM SMS CLASSIFICATION

A DISSERTATION

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE AWARD OF THE DEGREE
OF

MASTER OF TECHNOLOGY
IN
**COMPUTER SCIENCE AND ENGINEERING**

Submitted by

**SHWETA GUPTA**
**2K20/CSE/22**

Under the Supervision Of

**Dr. SHAILENDER KUMAR**
**(PROFESSOR)**

**DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING**

DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042

MAY, 2022

DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-11042

## CANDIDATE'S DECLARATION

I, (Shweta Gupta), Roll No: 2K20/CSE/22 student of M.Tech (Computer Science & Engineering), hereby declare that the Project Dissertation titled "DIMENSION REDUCTION FOR SPAM SMS CLASSIFICATION" which is submitted by me to the Department of Computer Science and Engineering, Delhi Technological University (formerly DCE), Delhi in partial fulfilment for the award of the degree Master of Technology, is original and not copied from any source without proper citation. This work has not previously formed the basis for the award of any Degree, Diploma Associateship, Fellowship or other similar title or recognition.

Place: Delhi (Shweta Gupta)
Date: 2K20/CSE/22

**DEPARTMENT OF**
**COMPUTER SCIENCE & ENGINEERING**

DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-11042

## CERTIFICATE

I hereby certify that the Project Dissertation titled "DIMENSION REDUCTION FOR SPAM SMS CLASSIFICATION" which is submitted by Shweta Gupta, Roll No: 2K20/CSE/22. DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING, Delhi Technological University, Delhi in partial fulfilment of the requirement for the award of the degree Master of Technology, is a record of the project work carried out by the student under my supervision. To the best of my knowledge this work have not been submitted in part or full for any Degree or Diploma to this University or elsewhere.

Place : Delhi

Date:

**DR. SHAILENDER KUMAR**
**(PROFESSOR)**
**SUPERVISOR**

# ACKNOWLEDGEMENT

I am very grateful to **DR. SHAILENDER KUMAR** (Professor, Department of Computer Science and Engineering) and all faculty members of Department of Computer Science and Engineering at DELHI TECHNOLOGICAL UNIVERSITY for their huge support and assistance for the completion of the project.

I would also like to acknowledge the much crucial role of university for contributing us with good equipped laboratories, infrastructure, testing facilities and positive space which motivates me to do work without any stumbling.
And I am also thankful to the laboratory staff, seniors who guided me and further to the colleagues who helped me in clarifying certain issues.

Shweta Gupta

(2K20/CSE/22)

# ABSTRACT

The invent of technology has given us a lot of outcomes. With the availability of network or internet a lot of change had been brought to the life of people. Earlier we were all depending on the physical modes of communication, commodities and interactions. But nowadays, we are all equipped with online or virtual world. We have online friends, online transactions, markets, finances and everything. Social Media and Mobile devices are one of the after-effect that were introduced with all these emerging technologies. A lot of data is generated every second from various online platforms. These data are highly valuable for business purpose to numerous organizations that are continuously studying this data.

SMS( Short Service Message) it is a text communication system where a person use to share information, data, news and communicate with his friend, family or other professional chats. As we know that with there is huge growth in the demand of mobile devices almost every person either from different financial background, of different age groups, belonging to different social communities, associated with different geographical areas or indulge in different economic activities owns these gadgets. Most of the population is owing Mobile phones nowadays. It had been found that the SMS or native messaging platform is one of the most preferred way of communication by different groups of people or organizations. With this the person can easily communicate and acquire knowledge he desires. But with some positive sides it had been ruined with some malevolent people or organization who tries to spoil the network by sending the rogue or undesired data to the recepients. These unsolicited SMS messages are termed as SPAM.

As we are aware that the mobile phones are having limited memory and filling that with such malicious data is not something we want.  Also, these messages that are received in bulk to the person may annoy him/her and may lead to skip of any urgent message that needs some earliest response or actions. A lot of Spam filtering tools are developed by numerous organizations but still the spammers had found some way to break the security. The spammers used to send the bulk of messages that tries to fill the user's inbox, some of the messages include promotional advertisements, some include fake offers that ask the user to reveal his personal information that would be used by the them to provide some financial rewards or some interesting vouchers, some may include the malicious links that may lead to the sites or software that would steal all private data of the user to the spammer by which he can make the financial loss to the person or the organization or some critical information loss. The spammers earns a good amout of financials through these activities. So we need to eliminate all such unwanted data from the network.

Here, in this work we had proposed an ensemble feature selection algorithm for the SPAM SMS classification that would help in classification of all such unsolicited messages from the platform. As there are large number of messages that are present in the database, we instead to manually doing such elimination move towards some automated ways. We usually employ machine learning algorithms for the task. But due to the presence of large number of features the machine learning models may sometime be guided by mischievous features. The large number of features may also lead to the high computation cost. So, to remove such redundant or irrelevant features from the dataset we employ feature selection algorithm. The Proposed feature selection algorithm’s performance had been tested against 5 chosen feature selection algorithms. Also the obtained dataset from all the feature selection algorithm is fed to to the three machine learning models. The results obtained were compared with the one deep learning model too.

# CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# LIST OF ABBREVIATIONS

| | |
|---|---|
| **NLP** | Natural Language Processing |
| **SMS** | Short Message Service |
| **RNN** | Recurrent Neural Network |
| **BOW** | Bag Of Words |
| **LSTM** | Long Short Term Memory |
| **SVM** | Support Vector Machine |
| **LR** | Logistic Regression |
| **XGB** | Extreme Gradient Boosting |
| **PSO** | Particle Swarm Optimization |
| **UCI** | UC Irvine |
| **RFECV** | Recursive Feature Elimination |

# CHAPTER-1

# INTRODUCTION

## 1.1 NATURAL LANGUAGE PROCESSING

NLP stands for Natural Language Processing, it is a field in Artificial Intelligence that has gain much recognition in last decades. With all new companies and institutes that are investing in Artificial Intelligence that had been in boom in all industries from last 15 years, NLP its subfield has been in predominantly existing with introduction of many new applications that will made the computer and human bond more friendly. It tries to allude the relationship between human and computer system via Natural Language or the language that humans speaks, writes, understands that would lead to more intelligent communication between human and machines. It incorporates empowering machines to control, dissect, decipher just as humans do. Natural language processing mainly regarded as two step process:

- Natural Language generation (NLG)
- Natural Language Understanding (NLU)

**Natural Language Generation:**

This field inculcates the production of natural language as output from numerous applications in an automated way.

**Natural Language Understanding:**

This process involves the interpretation of linguistics either present in text format, audio or video formats. The natural language understanding is found to be more hard than natural language generation because of some implicit nature of human language like coreference, sarcasm, Irony, homonyms, similar words with same meaning and many more.

There are numerous applications of NLP like sentiment analysis, speech recognition, coreference resolution, text classification, document classification, Named Entity Recognition and others. Here, in this work we had studied the impact of NLP for managing the social media platforms performing text categorization. Here we are mainly concentrating over the text data present over the SMS platform or generated by users over the platform.

## 1.2 SMS AND ITS FAQ'S

As with commencement of all new technologies and computer era now we are heading towards a virtual world people used to interact online rather than physical connections, Today the well-liked and immense relishness of different social media sites by almost every group of people either different age, economies, financially distinguished, or from different societies is something no one could deny. With continuous evolution of these researches that are being done each day and availability of internet the biggest network is available to all the people in such reduced price to almost every device that is making us move from offline to such larger virtual world. This virtual world can be accessed by all through their desktops, laptops, tablets, embedded devices and other available remote and mobile smart devices.

Social media is one of the most flourishing application. We have many social media platforms that are lurking over the network. Some most popular in the worldwide are: Facebook, Instagram, Twitter, WhatsApp,E-mail, SMS, YouTube, Snapchat and many more. We can connect through various applications or platforms that are lurking over the internet to our beloved ones, friends, family, business persons or for any other purpose. All these social media platforms, certain communities, groups enables us to interact and communicate removing distance barriers. Nowadays, these platforms are not affecting just the social or personal life but they had largely influenced the markets, instituitions and others sectors. Some statistics estimated the average usage of internet by a person to be 2 hours and 27 minutes[1]. Also some demographic reports revealed that there are almost 4.65 billion people that are utilizing social media platforms[2]. Figure 1.1 provides the past and estimated future statistics regarding the count of people utilizing social media sites.

As with heavy usage of different platforms a huge amount of data is generated over the network every second that is useful for many industries to understand and study the human behavior and for other multipurposes. This data is exploited for many researches and by several organizations.

[1] https://www.smartinsights.com/social-media-marketing/social-media-strategy/new-global-social-media-research/

[2] https://www.statista.com/topics/1164/social-networks/

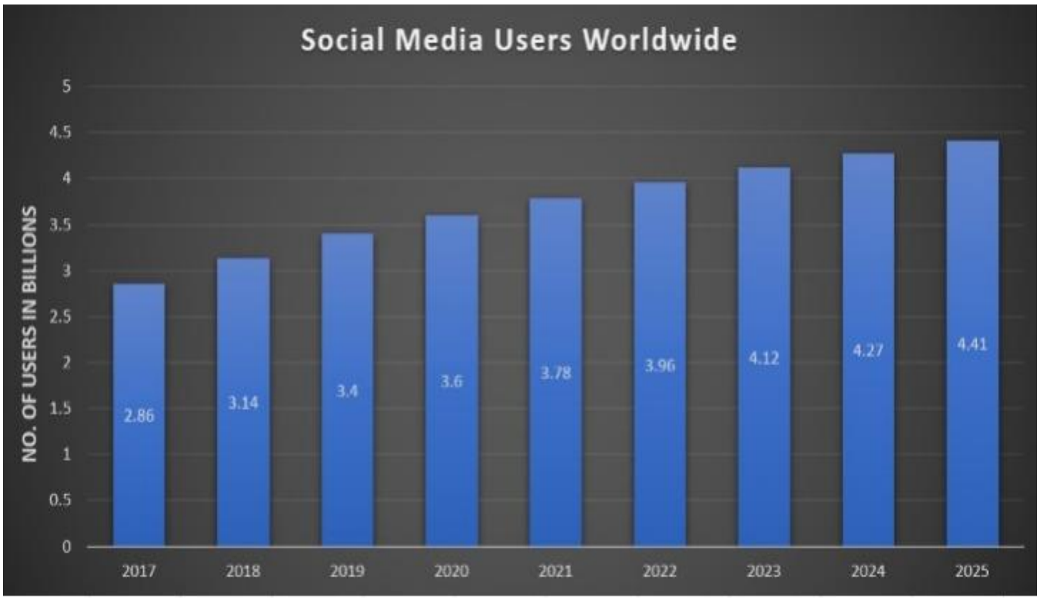


Figure.1.1. Statistics of Social Media users.

As we have discussed all the advantages that Social media has introduced to our life but it had also lead to some malicious activities over the network like clickjacking, fake news, threats, impersonisation, phishing, theft of data, abuse, cyberbullying, viruses. All these tries to create some rogue data over the platforms that make them vulnerable to heavy loss of data either for the organization or for person. All such activities and malevolent users are put under the name SPAM. The messages that are unsolicited are termed as "SPAM" whereas the others or solicited are termed as "HAM". The main categories of Spam are enlisted by the Figure 1.2. In this work we have done a brief study of SMS(Short Message Service) spam.

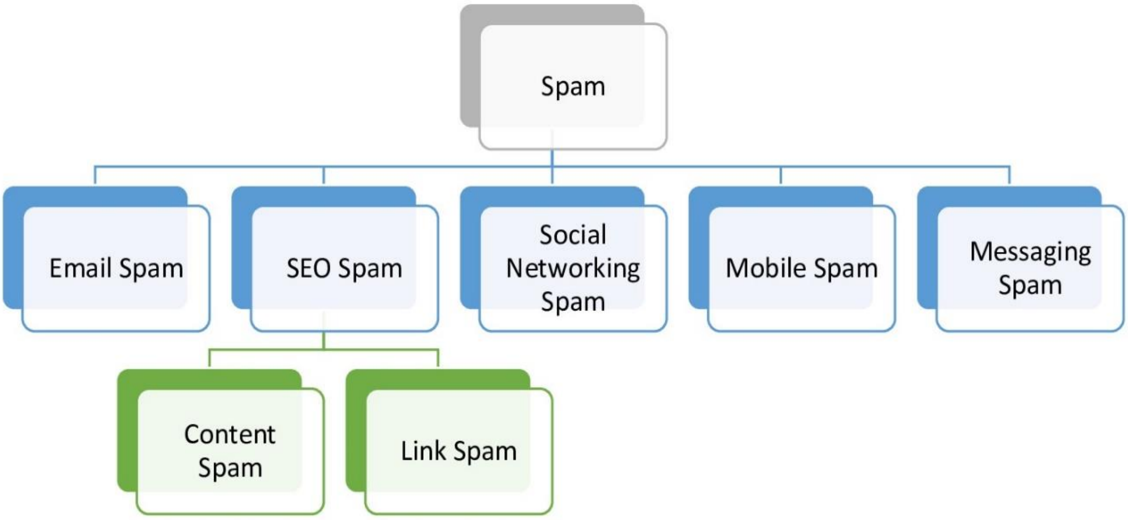


Figure 1.2: Some broad categories of Spam.

Native messaging applications are being preferred by most of the people for communication though with the presence of large number of social media platforms. These can be used for a personal, organizational, institutional, or for other promotional purposes. Among 1.2 billion mobile phone users in India around 750 million owns smarts phones[3]. Although it was also found that on an average the people choose to more frequently look after their SMS application rather than E-mail platforms. Over 18.7 billion SMS are communicated over the network for various purposes in one day[4]. As such large number of SMS that are initiated everyday through people, organizations they need to be managed and maintained by some automated mechanism. These SMS data is also induced with some malicious data or links that makes it spiteful. People tend to receive a large number of SMS that are undesired and unwanted termed as "SPAM SMS". These may comprise threat messages to a person, some links that may steal the private data or identity of person, some virus embed in the SMS that would make the phone hang or permanently damaged, some promotional messages, advertisements from different brands, industries where the person not wants to involved, some fake messages asking the user his OTP or other personal information to claim  fake offers or gifts by the person.  Such unwanted messages besides filling the user's phone space they may also lead to missing of some useful SMS that would need some instantaneous action or response. Figure 1.3 shows the acceptance of native messaging platform.

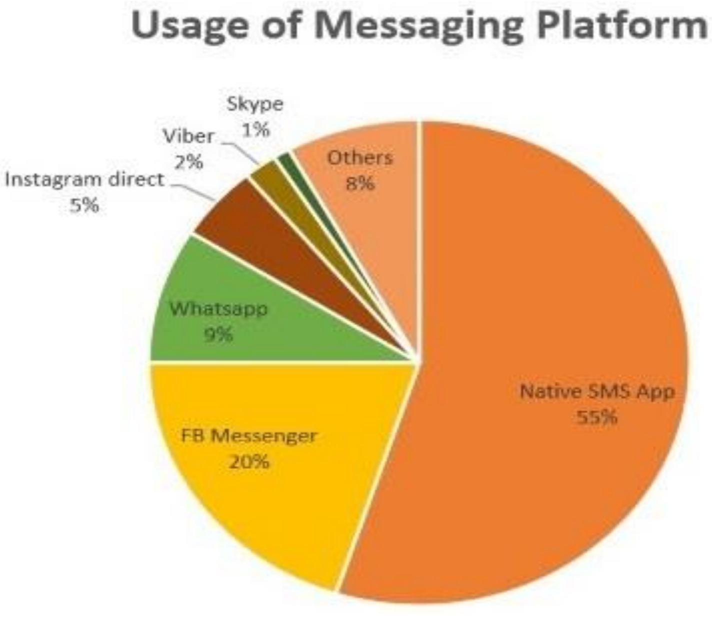


Figure 1.3: Usage of Different Messaging Platform

[3] https://www.business-standard.com/article/current-affairs/india-to-have-1-billion-smartphone-users-by-2026-deloitte-report-122022200996_1.html

[4] https://shso.vermont.gov/sites/ghsp/files/documents/Worldwide%20Texting%20Statistics.pdf

From above it is clear that native messaging platform is the most preferred platform by the people for communication. Table 1.1 shows the key points providing the acceptance of marketting organizations for SMS over different other available Social Media platforms.

Table 1.1 Comparison of SMS with other Social media websites for marketing[5]

| **Strategies** | **Description** | **Response** |
|---|---|---|
| **REACHING TARGET AUDIENCE** | Sending Marketing SMS to reach intended targets | SMS is preferred than any other social media platform |
| **ENGAGING CUSTOMERS** | To make clients aware of new products or services | Higher click rates are found over SMS platform than any other social media platform.<br>98% of SMS are opened within 3 minutes |
| **CONVERSION RATES** | Inclusion of more audience over the platform either through signup, sales or etc. | SMS have given a higher marketing conversion rate of about 47% greater than any other social media platform. |
| **MARKET PENETRATION** | With how much easiness these platforms can help you enter the markets | SMS is found to provide more easier method to penetrate into the market for new comers |

In this work a brief analysis of how different existing social media sites are affecting the world and inclusion of SPAM that makes this network venomous is performed.

From all the above discussion it can be concluded that SMS is an effective platform that is utilized by large number of people. The removal of SPAM is some mandatory and essential task that had to be performed. We had tried to eliminate the SPAM SMS from the platform. As we know that SMS is an text that could be of any length so deploying all such big SMS to the classification models sometimes lead to wrong results due to the presence of redundant features or some irrelevant features. So, we had included an additional step where we tried to remove the less significant features from the dataset also termed as “Feature Reduction” or we can say selecting the most significant features from the dataset that would be most beneficial for classification and would also be able to correspond the original dataset i.e. “Feature Selection”. We proposed a novel ensemble feature selection algorithm that is compared with other 5 feature selection algorithms : Recursive Feature Selection with SVC as classifier, Recursive Feature Selection algorithm with Random Forest as classifier, Low Variance, Particle Swarm Optimization, Mutual Information, Chi-Square. The proposed dataset that is obtained from each of the feature selection algorithms is deployed over three classification models: Logistic Regression, Support Vector Machine and Extreme Gradient Boosting. The proposed work was also found to be relatable to deep learning model LSTM.

[5] https://blog.textedly.com/sms-marketing-vs-social-media-marketing

# CHAPTER 2

# DATASET DESCRIPTION

All experiments and study has been developed over the dataset that had been taken from UCI MACHINE LEARNING REPOSITORY.

**UCI MACHINE LEARNING REPOSITORY**:

It is public repository that was initiated by David Aha along with his fellow mates at UC Irvine in1987. It is receiving its funds from National Science Foundtion. It has a good database entailing quality datasets that can be exploited by researchers, educational institutes, organizations, students and educators all over worldwide. It has been found as one of the top most cited library for the research.

The dataset selected for this work is “SMS SPAM”dataset[1]. The dataset is a CSV(Comma separated file) having 5574 total instances. The following Table.2.1 represents the description of the given dataset.

Table 2.1: Description of Dataset evolved

| | | | | | |
|---|---|---|---|---|---|
| **Data set Characteristics:** | Multivariate,Text, Domain-Theory | **Number of instances** | 5574 | **Area:** | **Computer** |
| **Attribute Characteristics:** | Real | **Number of Attributes** | N/A | **Date Donated** | **2012-06-22** |
| **Associated Tasks:** | Classification,Clustering | **Missing Values?** | N/A | **Number of Web Hits** | **387064** |

The pictorial representation of the statistics regarding the number of instance of SPAM or HAM messages in the opted dataset is given by Figure 2.1 whereas the token statistics are presented by the Figure 2.2.

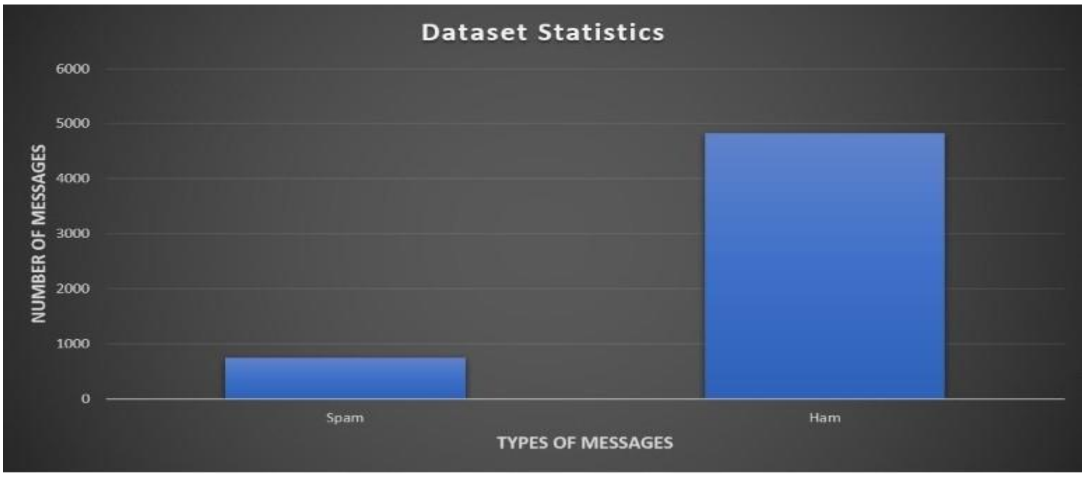


Figure 2.1 Statistics displaying dataset instances distribution.

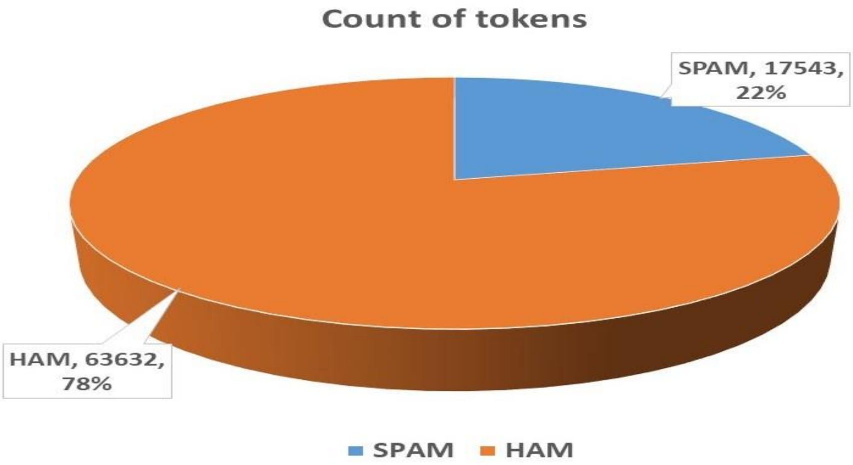
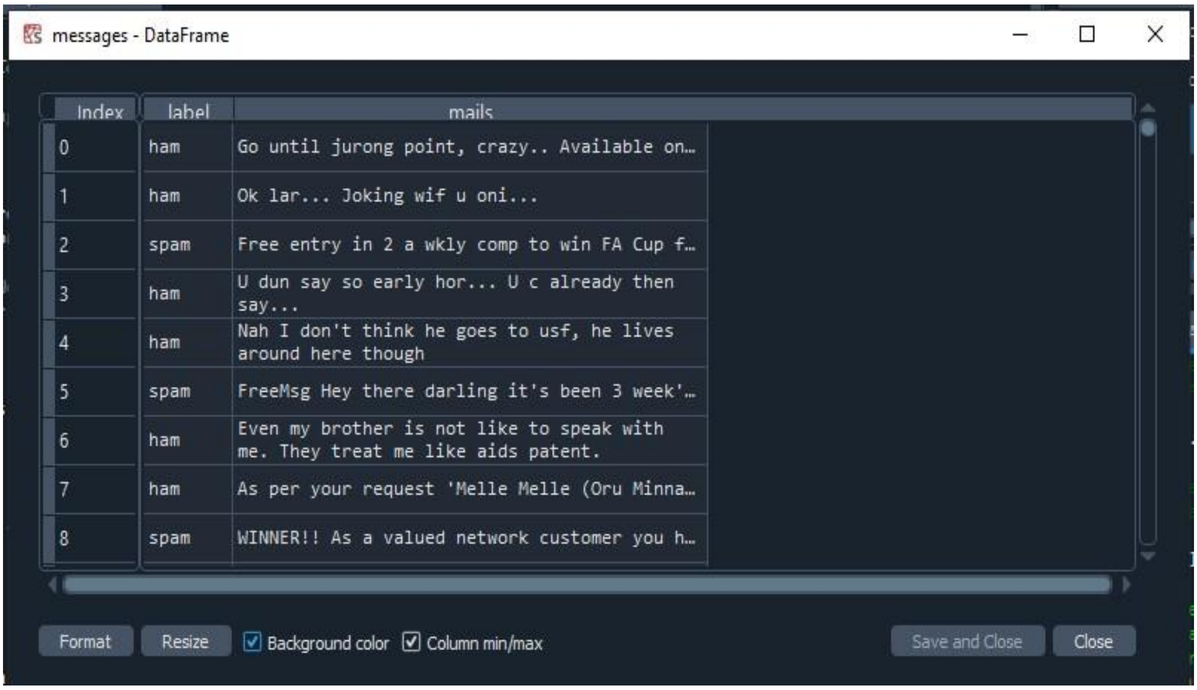


Figure 2.2 Token statistics for the opted dataset.

Figure 2.3 gives the pictorial view of some randomly chosen instances of SPAM and HAM messages.



| Index | label | mails |
|---|---|---|
| 0 | ham | Go until jurong point, crazy.. Available on… |
| 1 | ham | Ok lar... Joking wif u oni... |
| 2 | spam | Free entry in 2 a wkly comp to win FA Cup f… |
| 3 | ham | U dun say so early hor... U c already then say... |
| 4 | ham | Nah I don't think he goes to usf, he lives around here though |
| 5 | spam | FreeMsg Hey there darling it's been 3 week'… |
| 6 | ham | Even my brother is not like to speak with me. They treat me like aids patent. |
| 7 | ham | As per your request 'Melle Melle (Oru Minna… |
| 8 | spam | WINNER!! As a valued network customer you h… |

Figure 2.3: Sample dataframe of proposed Dataset

We can depict from the given characteristics and specifications of the dataset that it is an imbalanced dataset that can be related to the dataset that is prevailing in the environment. So that our model could be benefitted and utilized for some real world applications.

# CHAPTER 3

# DATA PREPROCESSING

This chapter provides the brief introduction of data processing techniques that are performed before deploying the dataset to classification or feature selection models.

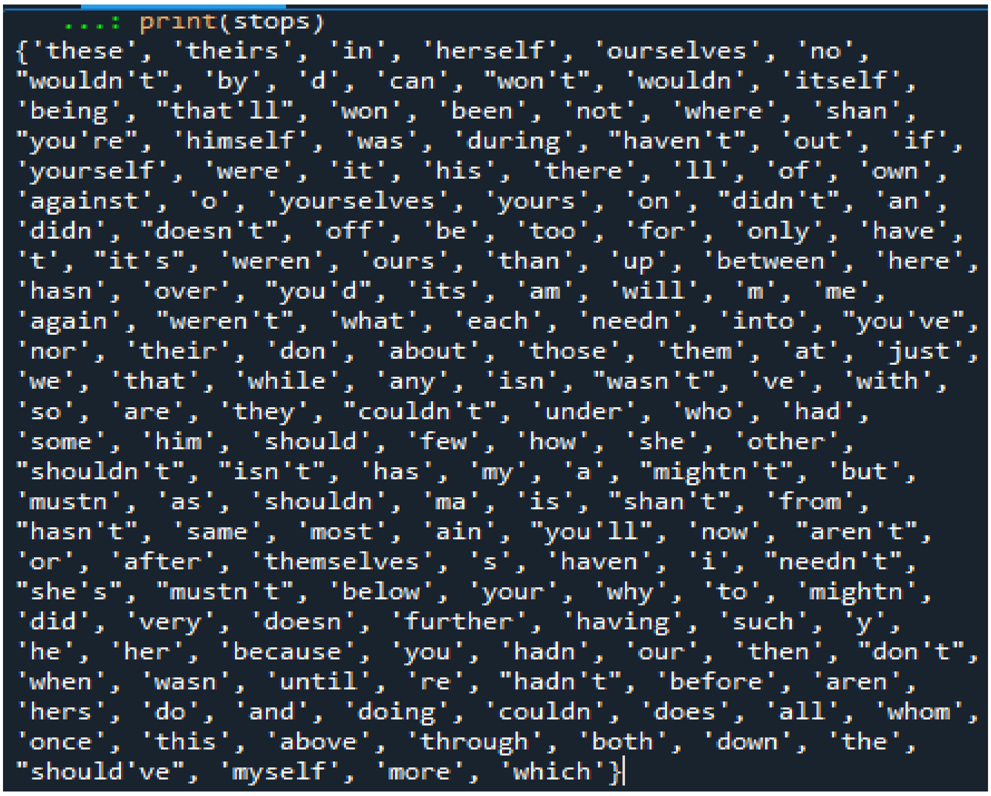
**NEED FOR DATA PREPROCESSING:**

Most of the data present over the internet is unorganized or semi-structured. The database may contain some missing values, some noisy data, inconsistent or duplicate data. The machine learning models may not be able to acquire such data properly. So some manipulation is done over such datasets before they are fed to the models. It also helps in enhancing the performance and accuracy of the machine learning models.

The adopted techniques in this work are as listed below:

1. **STOPWORDS REMOVAL:**
   Stopwords usually comprised of words that are utilized commonly. Most of the search engine programs are programmed to ignore these words. These are words like : “the”, “a”, “an”, “in”, “on”, “was”, “were”, “it”,….. . These are some commonly used words that are present in both the spam or unsolicited data and the ham or solicited data. So while classifying our data these could not provide much help as they are found in both the categories. Actually these words are not that much important to classify a message as spam or ham or these words do not convey much useful information that could be used for classification purpose so just to save our computation time and space and also to save from curse of dimensionality we remove all such words from our dataframe.

   NLTK( Natural Language Tool Kit) is a python library that has been deployed with stopwords of 16 languages. We had also utilized this library for stopwords removal from our dataset. The list of stopwords as provided in NLTK stopwords library is given by Figure 3.1.



```
   ...: print(stops)
{'these', 'theirs', 'in', 'herself', 'ourselves', 'no',
"wouldn't", 'by', 'd', 'can', "won't", 'wouldn', 'itself',
'being', "that'll", 'won', 'been', 'not', 'where', 'shan',
"you're", 'himself', 'was', 'during', "haven't", 'out', 'if',
'yourself', 'were', 'it', 'his', 'there', 'll', 'of', 'own',
'against', 'o', 'yourselves', 'yours', 'on', "didn't", 'an',
'didn', "doesn't", 'off', 'be', 'too', 'for', 'only', 'have',
't', "it's", 'weren', 'ours', 'than', 'up', 'between', 'here',
'hasn', 'over', "you'd", 'its', 'am', 'will', 'm', 'me',
'again', "weren't", 'what', 'each', 'needn', 'into', "you've",
'nor', 'their', 'don', 'about', 'those', 'them', 'at', 'just',
'we', 'that', 'while', 'any', 'isn', "wasn't", 've', 'with',
'so', 'are', 'they', "couldn't", 'under', 'who', 'had',
'some', 'him', 'should', 'few', 'how', 'she', 'other',
"shouldn't", "isn't", 'has', 'my', 'a', "mightn't", 'but',
'mustn', 'as', 'shouldn', 'ma', 'is', "shan't", 'from',
"hasn't", 'same', 'most', 'ain', "you'll", 'now', "aren't",
'or', 'after', 'themselves', 's', 'haven', 'i', "needn't",
"she's", "mustn't", 'below', 'your', 'why', 'to', 'mightn',
'did', 'very', 'doesn', 'further', 'having', 'such', 'y',
'he', 'her', 'because', 'you', 'hadn', 'our', 'then', "don't",
'when', 'wasn', 'until', 're', "hadn't", 'before', 'aren',
'hers', 'do', 'and', 'doing', 'couldn', 'does', 'all', 'whom',
'once', 'this', 'above', 'through', 'both', 'down', 'the',
"should've", 'myself', 'more', 'which'}
```

Figure 3.1. Stopwords list

2. **STEMMING:**

Mostly we use the different inflectional forms or derivationally related form of the words for some grammatical reasons or to enhance meaning of some words in the sentence. These words are like popular, popularly, popularity all these have some different spellings but inherently they all provide same meaning. So inspite of considering all these words as some different we want to make them similar by converting them to their base words like “popular” for above example. This is termed as “stemming”. Otherwise our classification model would treat them all as different and would lead to more features. Stemmer is a program that performs stemming. Many Stemmers are available some are listed below:

- Paice stemmer
- Porter Stemmer
- Lovins Stemmer

In our work we employed porter stemmer. It is an effective program developed in 1980. It is based on sequential algorithm that works in 5 phases.

3. **Bag Of Words (BOW):**

As we know how much efforts we would make but eventually the computer understands only one language ie binary so we need to make all the text to binary representation ie, we need to convert our words to some vector so for that here we will be utilizing BOW feature extraction algorithm.

After performing Stemming, the next step which we have to implement is Bag of Words. Basically, bag of words is a method of extracting features from text which further used in the machine learning models. In this approach, we take all the words in every message , then calculate the frequency of each word, we choose some specific words that appears oftenly more number of times than other words. The BAG OF WORD model is an improving on portrayal utilized in regular language handling and data recovery.

In this model the data is addressed as a bag (multiset) of its words, disregarding grammar or rules or even words but keeping multiplicity. It is also being used for computer vision or for pattern recognition field.

The BAG OF WORD model is normally utilized for document classification. Here the occurrence of each word is regarded as a component for preparing a classifier. The frequency of the word is given importance by this model it does not give any relevance to the order of the words.

# CHAPTER 4

# FEATURE SELECTION

The SMS may be composed of long characters, or large number of words. These large messages may turn into a dataset of large number of features which will be hard to learn by the predictive model. Feature selection is the process of lowering such high dimensional dataset to some corresponding lower dimension dataset that could be utilized by the classification model to give more accurate result with cost effective experiments. It is regarded as necessitate step for text categorization[2].

The features that are present in the dataset can be classified in three broad categories:

- **Redundant Features:** These are the features that may be already present in the dataset in some other format or their meaning had been conveyed to the classification models with some more strong features.
- **Relevant Features:** These are the features that are relevant to the classification models. or these are useful to the classification model to categorize the data.
- **Irrelevant Features:** These are the features that try to misguide the classification models or convey some wrong information to the models that will lead to wrong classification of instances by the model.

Feature selection is also regarded as process have multiple goals. The aim of feature selection algorithm is not only limited to elimination of undesired or irrelevant features but at same time it wants to withdraw redundant features. Hence, it can be assumed to be selection of the features that have more association with the labels of categories and less similarity with others features that are present in the feature set.

Figure.4.1 gives the pictorial representation of key steps involved in the feature selection algorithms.

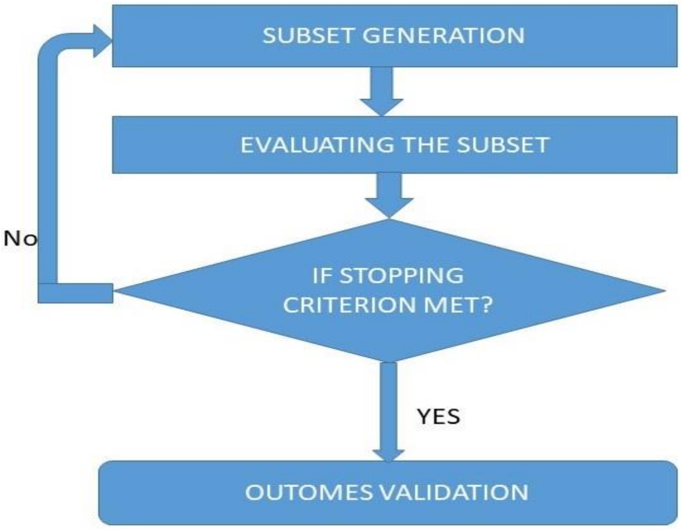


Figure4.1. Feature selection process

## 4.1 CATEGORIZATION OF FEATURE SELECTION PROCESS

The supervised feature selection techniques are classified in three broad categories:

- Filter Based
- Wrapper Based
- Intrinsic

## WRAPPER BASED FEATURE SELECTION TECHNIQUES:

Wrapper based feature selection techniques[4][5] are greedy approach that tries to fit the produced subset of dataset utilizing opted classification model against the selected evaluation function. They try to obtain the reduced most optimal features subset by iterative algorithm. Some examples of wrapper based feature selection algorithms are : forward feature selection algorithm, backward feature selection algorithm. Fig h gives the pictorial representation of Wrapper based feature selection methods.

**ADVANTAGES:**

- Works with combination of classifier
- Could find feature dependencies more comprehensively than filter based feature selection method
- Provides good generalization

**DISADVANTAGES:**

- Expensive
- More time complexity
- Prone to overfitting

## FILTER BASED FEATURE SELECTION TECHNIQUES:

Filter based feature selection techniques[6][7] are the statistical models that rank the features using some formulations. Either they try to rank on the basis of relevance with the target features or they may consider to rank on the basis of correlation present among the features set. Then using some threshold value the features satisfying that threshold value are accepted by the model. Some examples of filter based feature selection techniques are Euclidean distance, pearson correlation. Fig I gives the pictorial representation of filter based feature selection algorithms.

**ADVANTAGES:**

- Cost effective
- Less time complexity as compared to wrapper based algorithms
- Less prone to overfitting
- Good generalization
- Scalable to high dimenions

**DISADVANTAGES:**

- Do not interact with predictive model
- Considers features as independent sets and ignores dependency which lead to less productive feature subset selection.

## INTRINSIC FEATURE SELECTION TECHNIQUES:

These are also termed as Embedded or hybrid feature selection models[8][9] as they induce the effect of both wrapper and feature selection algorithms. Example: Lasso, Ridge.

**ADVANTAGES:**

- More accuracy
- Less training time complexity
- Consider dependency among features
- Less prone to overfitting

**DISADVANTAGES:**

- Costly
- Cannot be utilized efficiently for small feature subset.

## 4.2 BRIEF DESCRIPTION OF EMPLOYED FEATURE SELECTION TECHNIQUES

The feature selection algorithms(Recursive Feature Selection, Low Variance, Chi-Square, Mutual Information and Particle Swarm Optimization) adopted in this research study to compare the proposed ensemble feature selection algorithms are as given below:

1. **LOW VARIANCE (LV):**
   It is a filter based feature selection algorithm. It is a model based on the statistical concept that a feature that is present in the almost all the instances cannot help us to distinguish between the SPAM and HAM messages. It takes into account only the input features. It do not consider the target feature or output label. Here we choose randomly chosen value as threshold value. Any feature having value less than that is discarded by the model. Equation 4.1 provides the mathematical term for calculating the threshold value [10]:
   $$variance\ score = p(1-p) \qquad 4.1$$

2. **Chi-Square ($\chi^2$):**
   It is also a filter based feature selection algorithm. It considers the target value or the class label. It is based on the chi- square test which is a statistical test proposed by karl pearson in 1990. Here we select the n_ features that are having the highest correlation with the target value or output label. In this model the value adopted or the proportion of features “n” to be selected is given by us. It measures the relation among the stochastic variables throwing out the features that are not optimal or irrelevant. The mathematical term for the model is given by Equation 4.2 referenced from [10].
   $$\chi^2 = \frac{(observed\ frequency - expected\ frequency)^2}{expected\ frequecy} \qquad 4.2$$

3. **MUTUAL INFORMATION (MI):**
   It is filter based feature selection algorithm. It is based on information gain. It is a statistical techniques that tries to find out the strength of knowledge for one variable by studying other variable. This statistical technique not only considers the relativeness between the features of the feature set but also concentrate how those relevant features are useful for the target output class or label. Here, we calculate the score function for all instances given by Equation 4.3 [10]:
   $$score(X_i) = I(X_i\colon Y) - \beta \sum_{X_j \epsilon S} I(X_k\colon X_j) \qquad 4.3$$

   X$_i$ is the feature that is considered or whose value we are calculating.
   $I(X_i\colon Y)$ denotes the connection between the input and output variable.
   $\beta \sum_{X_j \epsilon S} I(X_k\colon X_j)$ represents the association among the other features and evaluated feature.

4. **RECURSIVE FEATURE ELIMINATION (RFE):**

Recursive Feature Elimination[11][12] is a wrapper based feature selection algorithm. We can employ the classification model either the one that is used for classification or any other for selecting the optimal feature subset. The classification models that could be engulfed in the RFE are : perceptron, SVM, Random Forest, Decision Trees. Here in this model we iteratively tries to find the optimal feature subset using the classification model against the employed or chosen performance metrics. It tries to find the most reduced feature subset but with higher performance score. It has an additional version that is RFECV( Recursive Feature Elimination with Cross Validation). In this we can do cross validation so that the classification model that is choosing the optimal subset of features not with just selected biased arrangement with other arrangements also.

5. **PARTICLE SWARM OPTIMIZATION (PSO):**

Particle swarm optimization is a heuristic algorithm that is based on the food searching observation of the animals. It was discovered in 1995 by kennedy & Eberhart [13][14]. Here the animal that is searching for food do not the exact location where the food is located but they could however estimate the distance they are far away from the food. Majorly the animals searches for food in groups or blocks. Each of the animals try to calculate the approximate distance between them and food, then the one having the least distance is followed by others. Here each animal remembers its own and block's current and best positions. After sometime or generation they again calculate these metrics. The algorithm is based on similar concept. Here in place of food it is lurking for best features as we had employed it here for feature selection purpose. Here after each iteration we would calculate the best optimal subset. Some of the hyperparameters that are being estimated in the algorithm are presented in mathematical form as given below.
For each iteration the value of $X_i$ and $V_i$ are updated as follows :

$$V_{id}^{t+1} = w * V_{id}^{t} + c_1 * r_1 * (p_{id} - X_{id}^{t}) + c_2 * r_2 * (p_{gd} - X_{id}^{t}) \qquad 4.4$$

$$X_{id}^{t+1} = X_{id}^{t} + V_{id}^{t+1} \qquad 4.5$$

Here both $X_i$ and $V_i$ are vectors of D-dimensions representing position and velocity. $P_{id}$ and $P_{gd}$ denotes the personal and global best values. $C_1$ and $C_2$ are learning constants. Its simulation for the feature subset selection is as follows, here the each animal is the possible solution. We start with some random chosen range or termed as population.The fitness function used in the pyswarm library for the discrete particle swarm optimization is as given by Equation 4.6.

$$f(X) = \propto (1 - P) + (1 - \propto)\left(1 - \frac{N_f}{N_t}\right) \qquad 4.6$$

In our study we had also utilized the pyswarm library [15][16] with most optimal choosen hyperparameters.

# CHAPTER 5

# CLASSIFICATION TECHNIQUES

The main aim of our study is to classify the given instance or SMS either as SPAM or HAM message. So we employ different variants or modified classification models for the task. Classification models can be broadly classified in two categories either as: Deep Learning Algorithm and Machine Learning Algorithm.

## 5.1 Deep Learning Algorithm:

In deep learning we actually tries to simulate how a human thinks or how a human brains works. The human body has many receivers that receives many signals or senses the environment like our eyes senses through sight, nose through smell, tongue through taste, ears through soundwaves prevailing in the environment, hands through touch. These are some receptors that accepts the signals from the surrounding to gather or retrieve the environment in their nearby.

With each signal received the human body tries to accept that and transfer to the brain for processing similarly, in the machines there are various devices that tries to sense, capture or understand the environment by various means and then they transfer the information retrieved to the CPU for processing. Here we tries to make our CPU learn certain algorithms that are based on the concept of neuron structure of brain. As brain have many thousands of neurons being ensemble in a network that helps it to take decision when there is any sense from the outer environment we introduces a similar structure in machines through the concept of "Perceptron" that will help machines to work like human brain.

It all started in 19's where many scientists proposed many ideas regarding this neural behaviour in machines that could simulate brain. Here we have utilized an LSTM deep learning model whose brief description is as given below.

### LSTM ( Long Short Term Memory):

Long Short term memory [17][18] is a deep learning architecture based on an artificial recurrent neural network (RNN). LSTM features feedback connection, unlike normal feedforward neural networks. It can handle not just individual data points(such as pictures), but also complete data streams (such as speech or video). LSTM may be used for task like unsegmented, linked handwriting recognition, speech recognition, or anomaly detection. Figure 5.1 gives the pictorial representation of Recurrent neural Network.

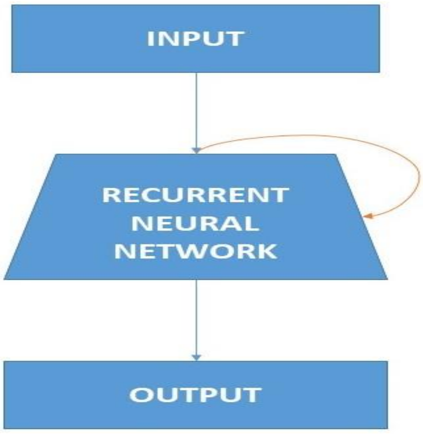


Figure 5.1 Recurrent Neural Network

An unfolded RNN would look similar to the one given by Figure 5.2.

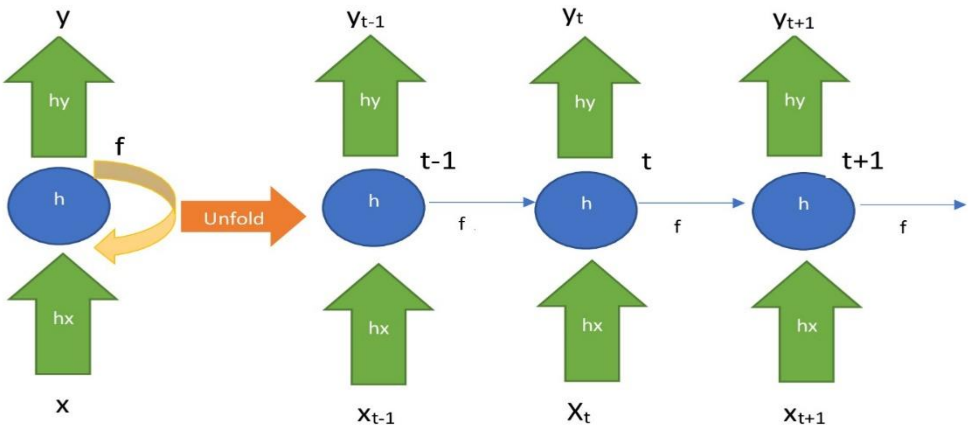


Figure 5.2 Unfolded RNN

LSTM is an artificial RNN which is designed to resolve the vanishing gradient problem which happens in RNN. LSTM tries to remember and forget things which is based on the context of the input. It is done by various gates which are used for different purposes. LSTM basically consists of four gates which are as follows:-

- Forget Gate
- Input Gate
- Input Modulation Gate
- Output Gate

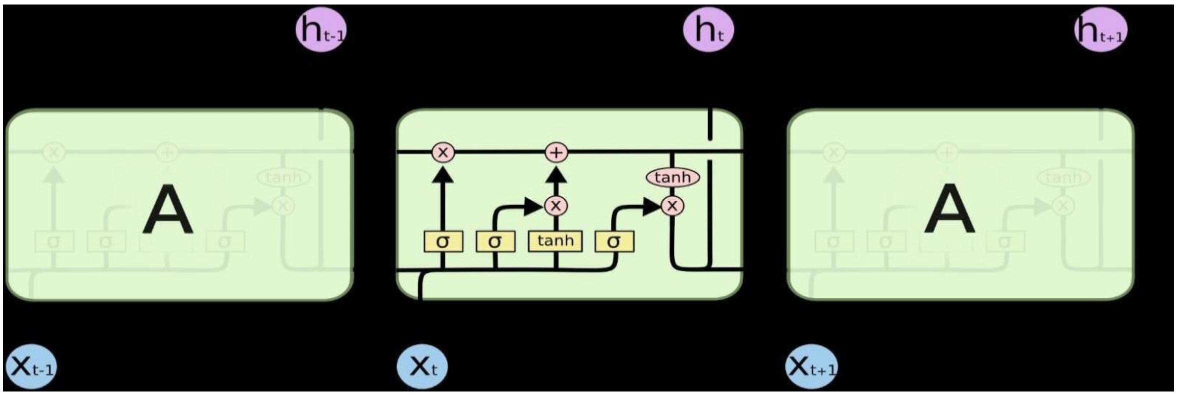


Figure 5.3 LSTM Network [19]

Figure 5.3 provides the pictorial representation of the LSTM model. The brief description how the processing is done is being described here. The initial step in LSTM is the forgetting of some words from the state of cell and it is done by the "Forget gate layer" picturized by figure 5.4.(i ) and its mathematical term is given by Equation 5.1. It focuses on inputs $h_{t-1}$ and $x_t$ , and gives output as the 0 or 1 in the state of cell i.e. $C_{t-1}$ . 0 means as "completely get rid of this" while 1 means as "completely keep this".

$$f_t = \sigma(W_t . [h_{t-1} , x_t] + b_t) \qquad 5.1$$

Fig L (ii) is depicting the Input gate whose function is keeping information in the state of the cell. It had been basically done in 2 parts i.e. a sigmoid layer and the "input gate layer". Equation (8) and Equation(9) gives the mathematical term for it. We have a "tan h" function which generates a vector for new values which will come further, $C_t$ which is extended to the state.

$$i_t = \sigma(W_i * [h_{t-1} , x_t] + b_i) \qquad 5.2$$

$$C_t = tanh(W_c * [h_{t-1} , x_t] + b_c) \quad 5.3$$

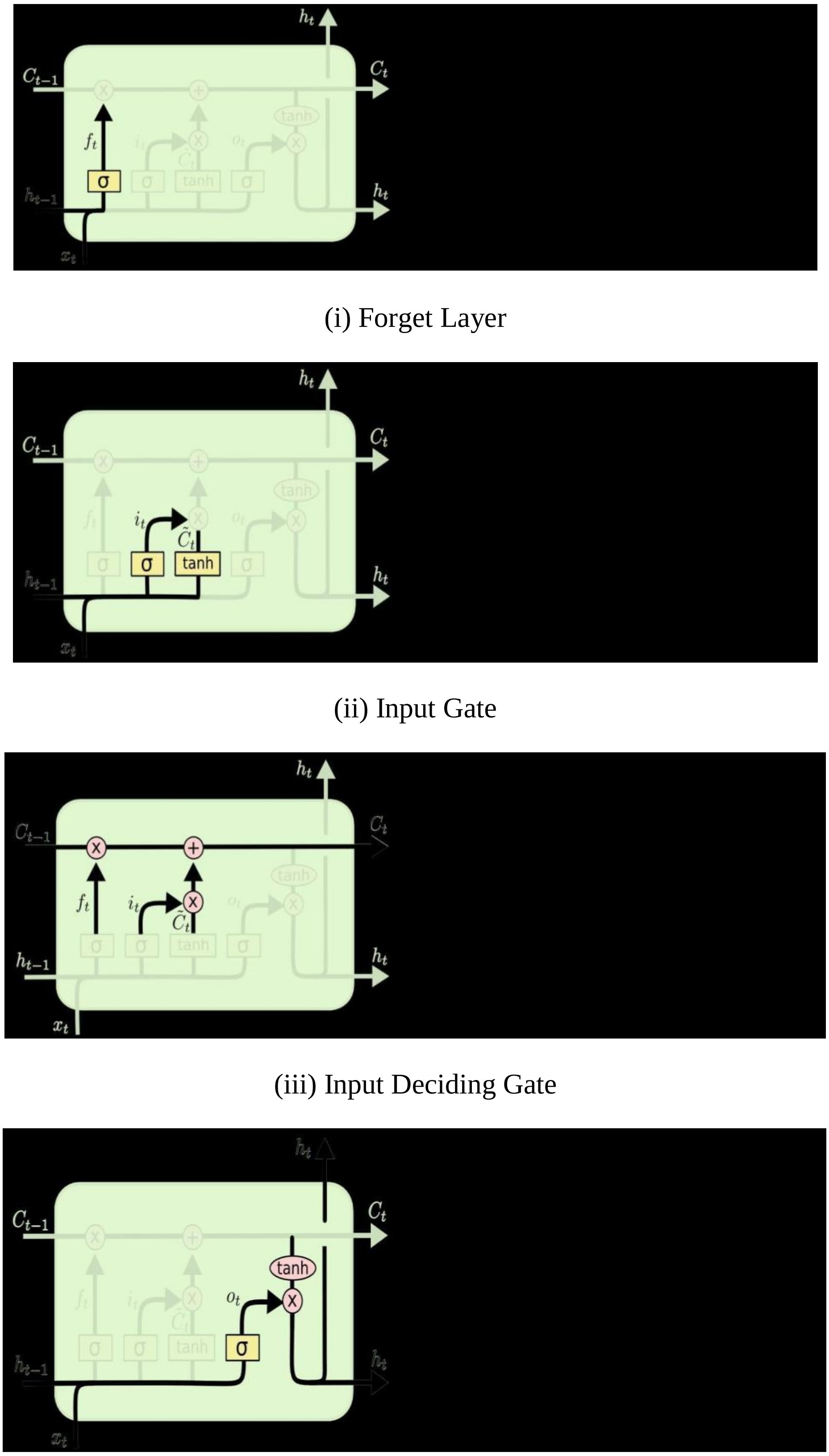
Now we add the old state of cell i.e. $\mathbf{C_{t-1}}$ added to the new state of cell i.e. $\mathbf{C_t}$ to conclude the input deciding gate. Whose mathematical term is given by Equation (10). And picturized by figL (iii). After that, we need to multiply old state by $\mathbf{f_t}$, and then add $\mathbf{i_t * C_t}$. So, the new values is a deciding factor to renovate each value of the state of a cell .

$$C_t = f_t * C_{t-1} + i_t * C_t \qquad 5.4$$

Now, we have to conclude the output which will be based on the state of cell. The mathematical calculation for output gate is given by Equation (11) and (12) and its pictorial illustration is given by Fig L (iv). So, intially we execute a sigmoid layer which decides the part of cell which gives the output then we pass the state of cell through "tan h" function and then it is multiplied to the output given by the sigmoid gate.

$$O_t = \sigma(W_o * [h_{t-1} , x_t] + b_0) \qquad 5.5$$

$$h_t = O_t \tanh(C_t) \qquad 5.6$$



(i) Forget Layer

(ii) Input Gate

(iii) Input Deciding Gate

(iv) Output Gate

Figure 5.4 Gates in LSTM network[19]

Some limitations of LSTM

- It takes too much time to train.
- Dropout is difficult to implement
- It is sensitive to different weight initializations.

Despite of certain limitations LSTM is in use for many applications and also various extensions and modifications of it also prevails.

## 5.2 MACHINE LEARNING ALGORITHMS:

Machine learning is a class of technology where we try to make machine automatically create models for data evaluation, just as we humans tries to do.

Machine learning algorithms are the engine of Artificial intelligence field , meaning it is the calculation that transforms an informational collection into a model. Which kind of model (supervised, unsupervised, regression, classification, etc.) should be used depends on the kind of problem when have been stucked into or for which we are deploying the model being proposed. , the machine capabilities, processor required, different resources required all are responsible and also the kind of data that had to be used for the purpose.
Here we have utilized three machine learning algorithms. Two are simple machine learning models and one is ensemble learning model.

### MACHINE LEARNING MODELS:

Here we employed two machine learning models: Support Vector Machine and Logistic Regression and one Ensemble model.

### SUPPORT VECTOR MACHINE

Though there are large number of machine learning algorithms that are prevailing in the environment but Support Vector Machine has gain more acceptance then other. This achievement could be raised as a result of outstanding results this algorithm has given in several fields. Support vector machine[20][21] are introduced as very effective machine learning algorithms that have its application in many fields comprising of regression, classification (it can be used for binary classification and also it can be extended to do multiclass classification though its implicity designed for binary or two class categorization problems). SVM algorithm utilizes the VC statistical theory.

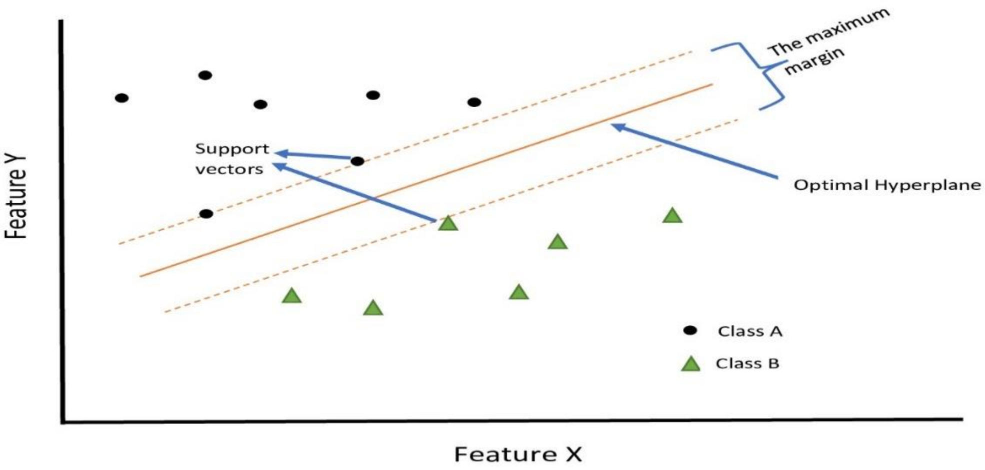
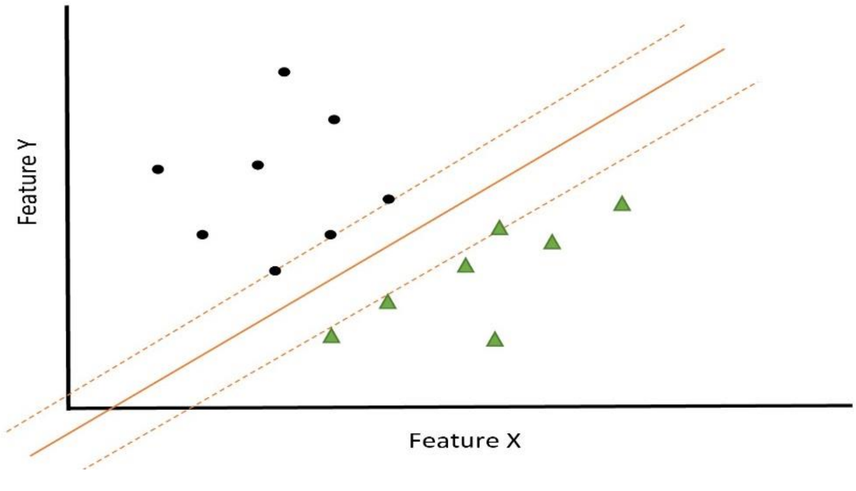
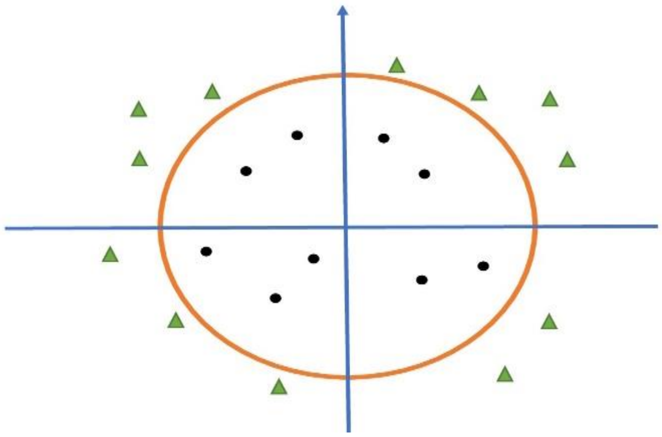


Figure 5.5 Support Vector Machine

The pictorial representation of Support vector machine is given by figure 5.5. Here, we can see that the there are two kind of points which are located on the graph these are simulating the two different kind of instances that are existing in the dataset which are needed to be classified as in our example they are SPAM or HAM mails. So here we plot the points on the graph. Figure.5.6 depicts two kinds of data here in Figure5.6.(i) the data is linearly separable but in Figure.5.6.(ii) the data is not linearly separable as here we cannot draw a simple line that could distinguish the data points in either categories.



(i). Linearly Separable Datapoints



(ii). Non-Linearly Separable Data

Figure 5.6. Distribution of data points

Here the line that distinguishes the data points of two different categories is termed as hyperplane. We can have more than one hyper plane we choose the one that can best differentiate the data points from each other belonging to different labels. The hyper planes are built by using support vectors. These are the point of instances that are present on the margin they helps us to find the extreme deviation that a model should consider. The hyperplane above the maximum margin hyperplane is termed as positive hyperplane and below one is termed as negative hyperplane. The mathematical equation for maximum margin hyperplane and positive and negative hyperplane is given by Equation 5.7,5.8 and 5.9.

$$g(X) = W^T X + b = \sum_{i \in SV} \alpha_i X_i^T X + b \qquad 5.7$$

$$W^T . X + b = -1 \qquad 5.8$$

$$W^T . X + b = 1 \qquad 5.9$$

Here we have other parameters also that help us to modify the algorithm according to our need. The most important function in this algorithm is “kernel function”. It helps in differentiating the data points that are not linearly separable by plotting those points to some higher dimensional space. Though it seems the process would induce more time complexity but by using some mathematical formulas the preferred way take less computational time. The function whose satisfaction make any function as kernel function is given by Equation 5.10 and its some popular variants are given by Table 5.1.

$$\sum_{i,j} k(x_i, x_j) c_i c_j \geq 0 \qquad 5.10$$

Table 5.1. Kernel Functions of SVM

| **Kernel Functions** | **Mathematical Formulations** |
|---|---|
| **Linear** | $k(x_i , x_j) = \ x_i . x_j$ |
| **Polynomial** | $k(x_i , x_j) = (1 + \ x_i . x_j)^{\mathrm{p}}$ |
| **Sigmoid** | $k(x_i , x_j) = \ e^{-\frac{\|\|x_i - x_j\|\|^2}{2\sigma^2}}$ |
| **Radial Basis Function** | $k(x_i , x_j) = \ \tanh(\beta_0 \, x_i . x_j + \beta_1)$ |

## LOGISTIC REGRESSION

Logistic regression [23][24][25] is an algorithm that has gained a lot of popularity. Thoughits name contains regression but there is an anomaly in its working as it is used for classification. The algorithm can be applied for both binary class and multi class classification. It had got its name as  logistic regression as it is based on linear regression principle, the algorithm that is utilized for regression problems. There are three different types of Logistic regression models that are categorized based on algorithm outcomes:

- Binomial Logistic Regression model
- Multinomial Logistic Regression model
- Ordinal Logistic Regression model

It is based on logistic mathematical equation which is given by Equation 5.11.

$$g(z) = \frac{1}{1 + e^{-z}} \qquad 5.11$$

Its derivation has possess a special property which it more popular and acceptable by the researchers. The mathematical term for its derivative is given by Equation 5.12.

$$g'(z) = g(z)(1 - g(z)) \qquad 5.12$$

It is a probabilistic statistical model. Here the probability for a given instance X to belong to a class is given by Equation 5.13

$$P\left({}^{t}/_{X}\right) = h(X)^{t}\ \left(1 - h(X)\right)^{1-t} \qquad 5.13$$

$$h(X) = g(\beta^{T}X) \qquad 5.14$$

$$L(\beta) = \ \prod\nolimits_{X} P(\ t/X\ ; \beta) \qquad 5.15$$

So, the maximal probable label or class is opted by the model. It uses logarithmic function which helps in reducing the time cost for the computation which makes the model more efficient. The mathematical terms for them are as given below.

$$complexity(\beta) = \ \sum\nolimits_{X} - t\log\bigl(h(X)\bigr) - (1-t)\log(1 - h(X)) \qquad 5.16$$

## ENSEMBLE LEARNING

Though we have many machine learning algorithms that are present today and their different variants. But still there are some limitation as when we are doing computation for some big dimensional data or some unevenly distributed data. The new concept of ensemble learning was introduced [29][30]. Here we compute the result as the aggregation of not just one algorithm or computational model but with a collection of algorithms or models. As here instead of just one model we are relying on numerous models for our analysis or results so it is more reliable. Here, the models can be combined or their results can be aggregated through two ways on the basis of which we categorise the Ensemble learning algorithms:

- Boosting algorithms
- Bagging algorithms

## BAGGING ALGORITHM

The name is derived from BOOTSTRAP AGGREGATION. Here we train the machine learning models over different samples of the dataset. Instead of feeding whole dataset to single model, we provide some randomly distributed subset of data to more than one instances of the opted machine learning model. Here each of the algorithm or model gets some distinguished or unique subset of dataset. Then we aggregate the results obtain using some statistical calculations. Figure 5.7 gives some pictorial working representation of the Bagging technique.

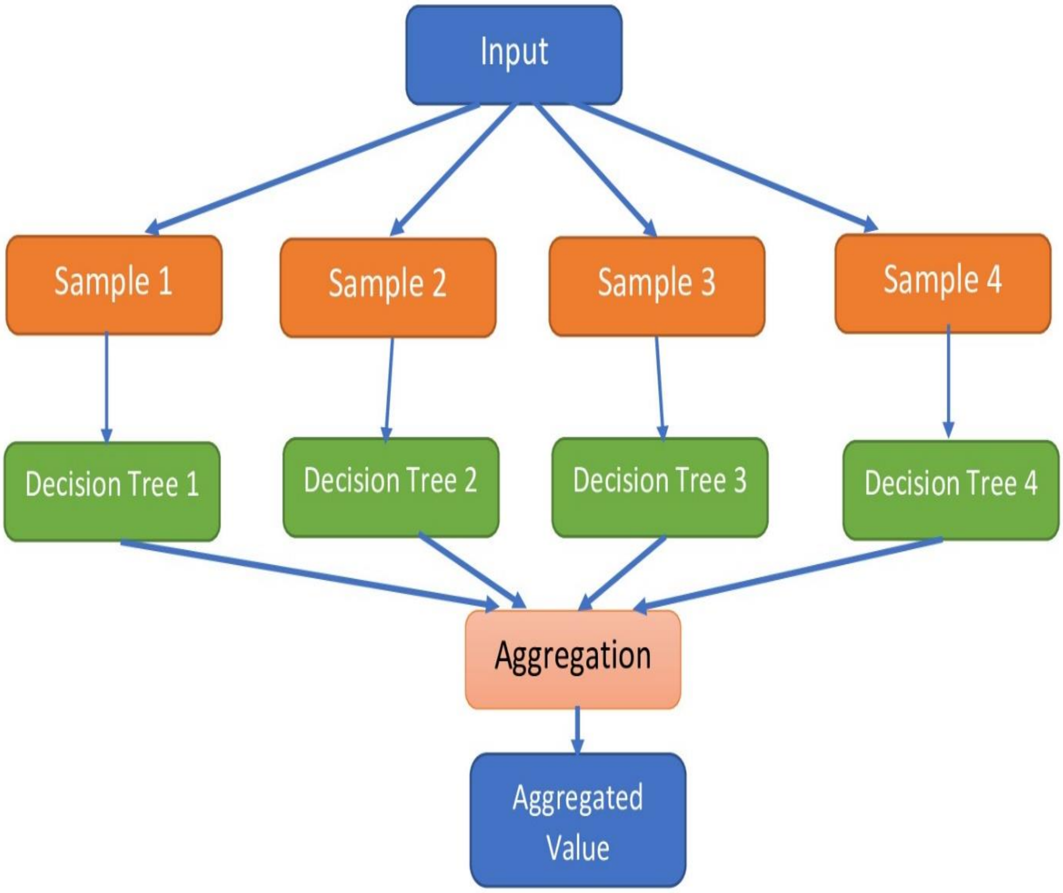


Figure 5.7 Bagging Algorithm

## BOOSTING ALGORITHM

Boosting ensemble learning is based on the fact to improve and work on the dataset that has been wrongly classified by the previous model. Here instead of feeing the model with the complete dataset or subset of data, the subsequent models are fed with the data that are misclassified by the previous ones. Here, we try to boost the learning of the previous models by this procedure. The outcomes of the models are aggregated by simple calculations involving averaging or voting. The efforts are made to make a strong model from the present weak models by training over the dataset instances that are hard to evaluate. Figure 5.8 gives the pictorial representation of the algorithm.

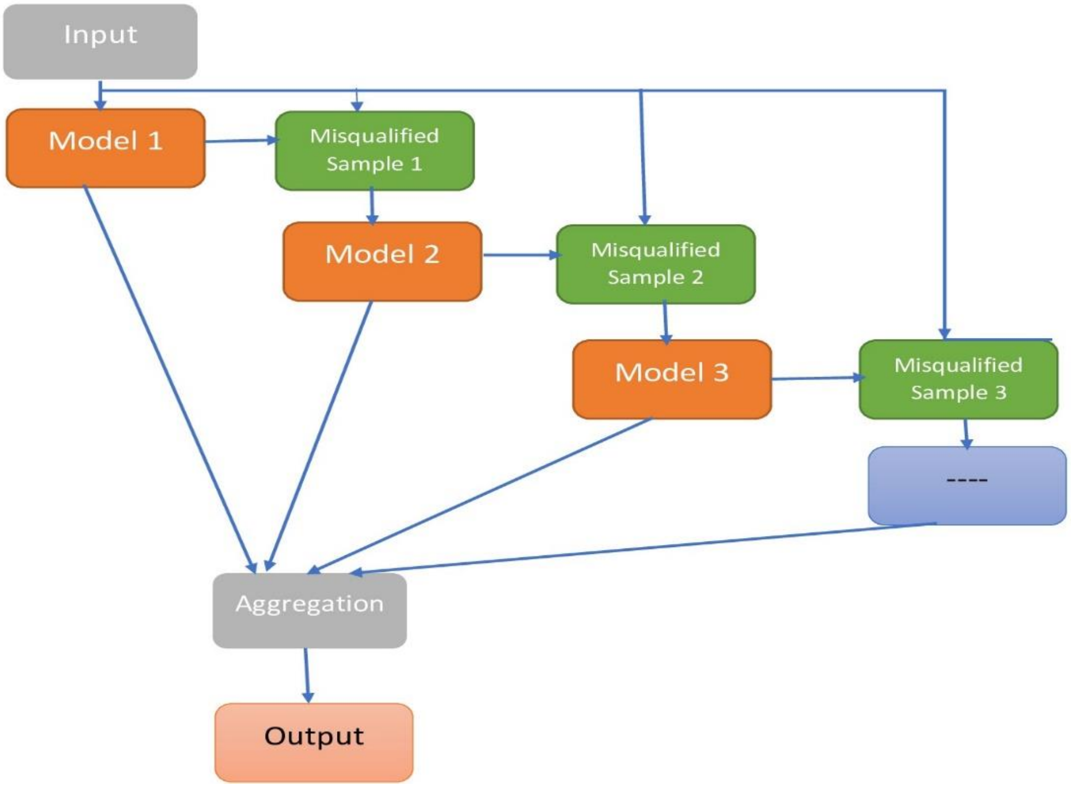


Figure.5.8 Boosting Algorithm

### Extreme Gradient Boosting (XGB)

Extreme Gradient Boosting[27][28] is an example of Boosting ensemble learning. This algorithm was put forward by Tianqui Chen, was elaborated in paper [26] in 2016. This model can be used for classification and regression problems. Here we use decision tree as the base models. We choose some hyper parameters tuned Decision tree model that we iteratively train for the given dataset for the misclassification done by the priori model. It is an optimized variant of Stochastic Gradient Boosting Algorithm. The model is highly scalable and it uses loglosss function for the error evaluation.

# CHAPTER 6

# PROPOSED MODEL

In this work a novel ensemble algorithm had been provided for feature selection. This algorithm is based on two level of filtering the features. It employs Recursive Feature Elimination algorithm using two different classification models that are fed into the wrapper algorithm. The algorithm uses two models that are: Support Vectors Machine and Random forest. The positive effects of the models are utilized in the ensemble algorithm by setting up the hyper parameters  as per the need of the model and field. The proposed algorithm removes the limitation of one algorithm by putting the another algorithm for the reference as the concept used by the boosting algorithms. Figure 6.1 gives the flow chart of the proposed algorithm.

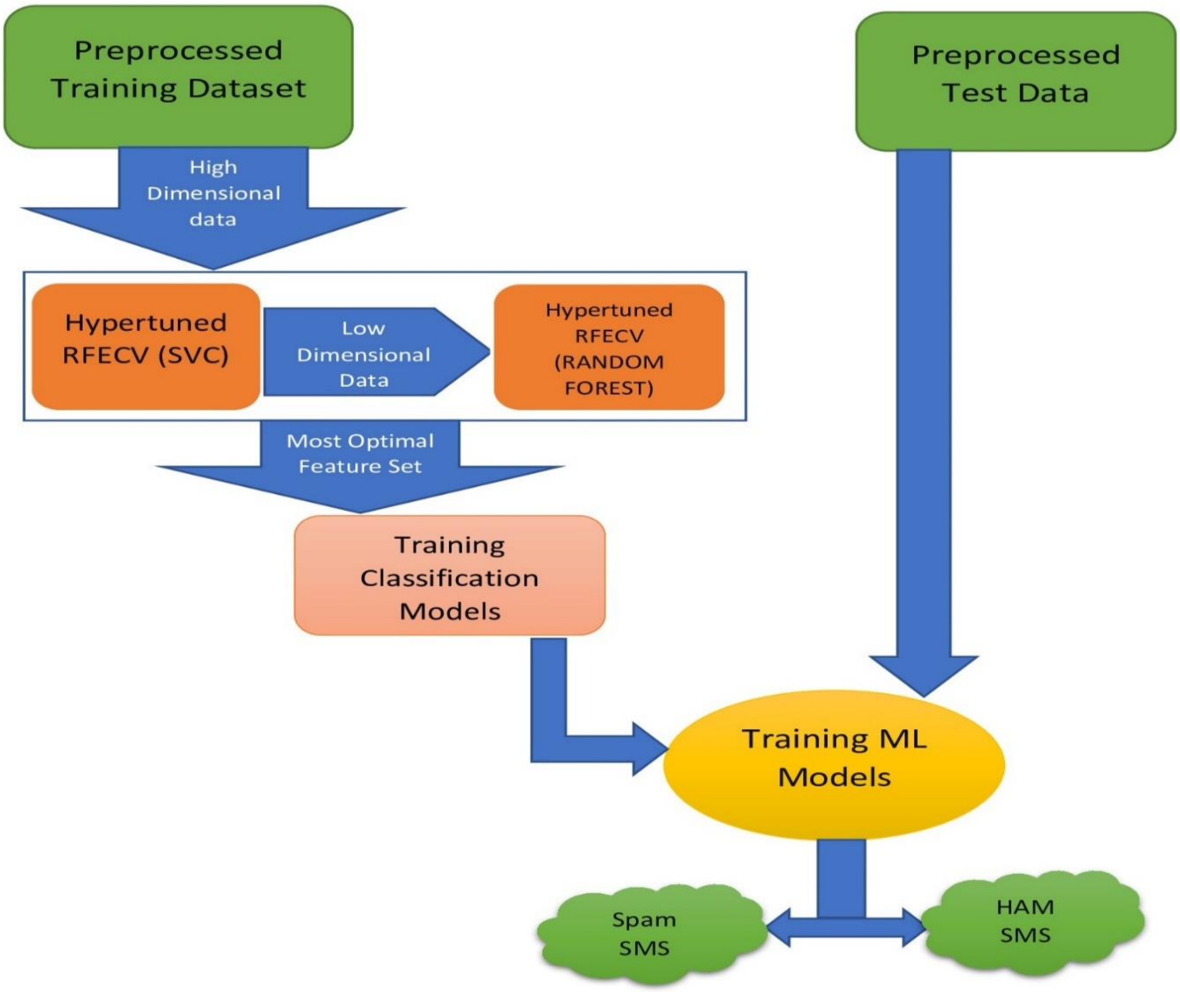


Figure 6.1 Flow chart of proposed algorithm.

The Proposed model also proves the effectiveness of standalone machine learning algorithms that can surpass even the neural networks that are too complex and unpredicatable. The support vector machine and the Random Forest algorithms that are employed in the proposed ensemble learning and fed to the RFE algorithm along with their parameters that enhance their performance are as given and explained in the below stepwise algorithm

The stepwise flow of work as proposed in this model is as given below:

**Step 1:** firstly, the dataset is selected

**Step 2:** As dataset may include some missing values or noisy data, the data preprocessing is done that includes tokenizing, stemming and stopwords removal.

**Step 3:** Then for feature extraction the Bag of Words algorithm is utilized

**Step 4:** For Feature selection we proposed the feature selection model based on boosting ensemble learning model.

Here, we first fed the data to the wrapper based Recursive feature Elimination algorithm employing Support Vector Machine. The hyper parameters were tuned as per the model requirement that could best classify the data.

Table 6.1. Parameterized Support vector and Recursive Feature Elimination wrapper class model is as given below:

| **SUPPORT VECTOR MACHINE** | | |
|---|---|---|
| **Parameters** | **Opted value** | **Reason** |
| **Kernel Function** | Linear | As the linear kernel is having the lowest complexity and is also find to be best classifier for text and also estimated with the results achieved on our dataset. |
| **Penalty Parameter** | 1 | It was tested for a wide range the best and optimal one was chosen against performance metrics |
| **RECURSIVE FEATURE ELIMINATION** | | |
| **Classification model** | SVC | As it's a light weight model and is scalable |
| **Step size** | 5% | ------------- |
| **validation** | Stratified cross validated | The stratified k fold is found to be more reliable |

| | | |
|---|---|---|
| | | than cross validation as here we can |
| **Performance metric** | accuracy | As our aim is to achieve reduced but more accurate and quality features |

**Step 5:** The results achieved from the above step are as follows:

Table 6.2. Achievement of RFECV with SVC

| **Attributes** | **Values** |
|---|---|
| **Number of features fed** | 6296 |
| **Number of feature retained** | 1590 |
| **Number of features removed** | 3986 |
| **Accuracy achieved** | 97% |
| **Percentage of features remained** | 25% |

**Step 6:** The retrieved reduced dataset is then fed to the hyper parameters set up Recursive Feature Elimination algorithm employing ensemble learning parameterized Random Forest as classification model. The hyperparameters opted for this model as per their demand are as listed below:

Table 6.3. Parameterized Random forest and Recursive Feature Elimination wrapper class model is as given below

| | **Random Forest** | |
|---|---|---|
| **Parameters** | **Opted value** | **Reason** |
| **Number of decision trees** | 50 | The range from 10 to 1000 was experimented and the best and lowest or the most optimal was selected. |
| **criterion** | entropy | Logarithmic function that has better performance than gini-index |
| | **RECURSIVE FEATURE ELIMINATION** | |
| **Classification model** | Random forest | As its an ensemble model that can learn many samples and give the aggregated and most optimal features subset |
| **Step size** | 1% | ------------- |
| **validation** | Stratified cross validated | The stratified k fold is found to be more reliable than cross validation as here we can |

|  |  |  |
|---|---|---|
| **Performance metric** | accuracy | As our aim is to achieve reduced but more accurate and quality features |

**Step 7:** The results achieved through above model are given as follows:

Table 6.4. Achievement of RFECV with ensemble model

|  |  |
|---|---|
| **Number of features fed** | 1590 |
| **Number of feature retained** | 845 |
| **Number of features removed** | 745 |
| **Accuracy achieved** | 100% |
| **Percentage of features remained** | 13% |

**Step 8:** The achieved reduced low dimension dataset is then fed to three machine learning models: Support vector Machine, Logistic Regression an Extreme Gradient Boosting.

**Step 9:** The achieved results that are obtained from the proposed ensemble feature selection algorithm are tested or evaluated against selected five feature selection algorithm: Recursive Feature Selection algorithm, Low Variance, Chi-Square, Mutual Information and Particle Swarm Optimization.

# CHAPTER 7

# EXPERIMENT AND RESULTS

The execution of series of experiments done in this work is described in this chapter. The efficiency of the proposed model is being proved by comparing the outcomes of the proposed feature selection algorithm with other five feature selection algorithm. The Feature set obtained from the feature selection algorithms and the proposed algorithms are fed to three machine learning models. The results are compared utilizing the performance metrics as illustrated further.

The overall structure being followed in this work is presented diagrammatically by the Figure 7.1.

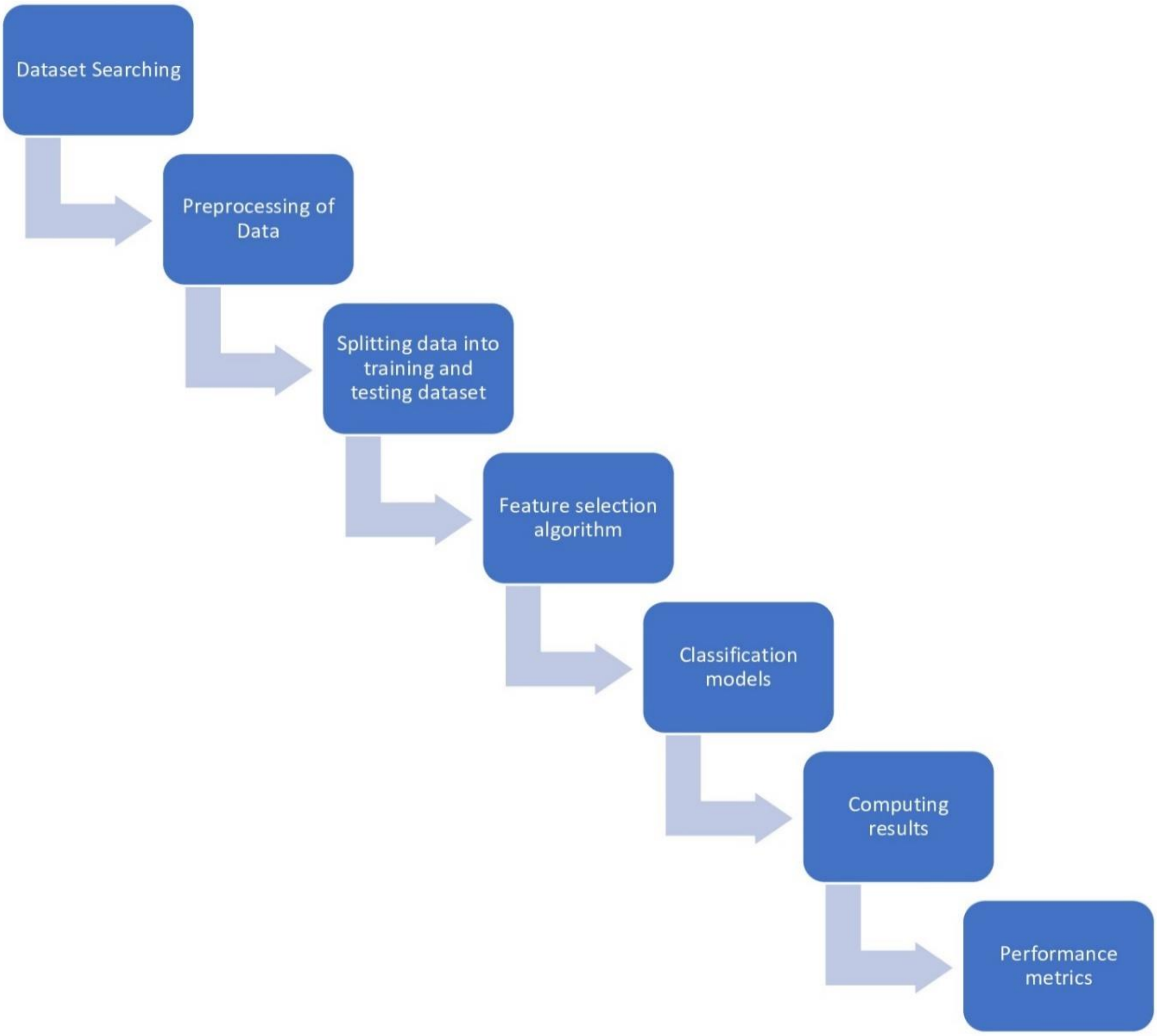


Figure 7.1 Experiments outline

The feature selection is the process of reducing the feature set to some optimal corresponding dataset that can replace the original dataset. The features obtained by using the proposed and other selected five feature selection algorithm is given by line graph presented Figure 7.2.

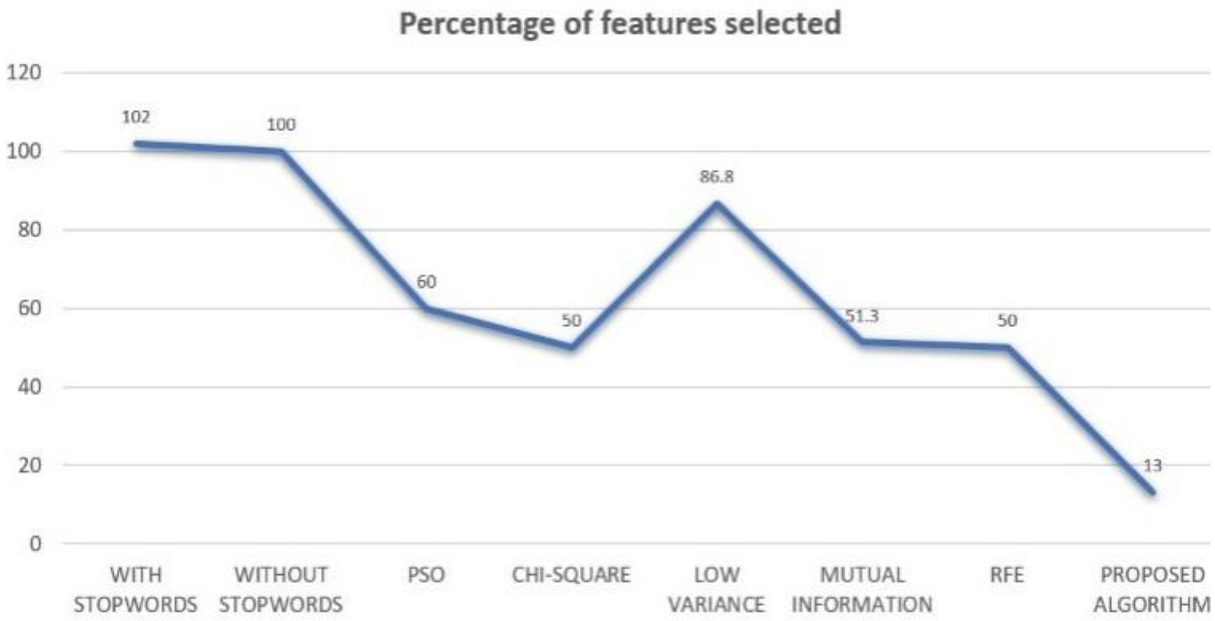


Figure 7.2 Percentage of features selected by feature selection algorithm

From the above figure it can be estimated that our proposed ensemble learning model is able to give the reduced dataset that is most optimal. It also shows that there are many redundant or irrelavent features that were present in the model which could not be removed by just single filtering algorithms as our model with RFE with SVM and Random forest alone cannot remove or filter out all the features that were not of that importance.

## PERFORMANCE METRICS

Performance metrics helps us to evaluate the results that were obtained from the different classification models. These used to compare the results obtained against the actual value for the test dataset.

The metrics utilized by us to evaluate the performance of our algorithm are : precision, recall and f-score.

$$precision = \frac{TP}{TP+FP} \qquad 7.1$$

$$recall = \frac{TP}{TP+FN} \qquad 7.2$$

$$f - score = \ (1 + \beta^2)\ \frac{Precision*Recall}{(\beta^2*Precision\ )+Recall} \qquad 7.3$$

Here, TP represents True Positive, TN represents True Negative, FP represents False Positive and FN represents false Negative.

The dataset distribution that is being used in this work is given by Table 7.1

Table 7.1 Dataset distribution statistics

| | **Percentage** |
|---|---|
| **Testing Dataset** | 75% |
| **Training Dataset** | 25% |

The results of the experiments is given by Table 7.2

Table 7.2 Experimental results

(i) SVM model with other feature selection algorithm

| Models | Precision | Recall | F-0.5 Score |
|---|---|---|---|
| **Without Feature selection** | 94.61 | 87.78 | 93.16 |
| **With stopwords** | 92.94 | 87.78 | 91.86 |
| **LV** | 95.53 | 92.43 | 94.89 |
| **MI** | 95.05 | 93.51 | 94.74 |
| **PSO** | 94.85 | 89.72 | 93.78 |
| **CHI** | 95.12 | 86.67 | 93.30 |
| **RFE(SVC)** | 95.53 | 92.43 | 94.89 |
| **RFE(RandomForest)** | 96.04 | 91.89 | 95.18 |
| **Proposed algorithm** | **98.27** | **92.43** | **97.04** |

(ii) LR model with other feature selection algorithm

| Models | Precision | Recall | F-0.5 Score |
|---|---|---|---|
| **Without Feature selection** | 94.47 | 85.56 | 92.54 |
| **With stopwords** | 93.97 | 86.67 | 92.41 |
| **LV** | 98.18 | 87.56 | 95.85 |
| **MI** | 97.54 | 85.94 | 94.98 |
| **PSO** | 98.13 | 85.40 | 95.29 |
| **CHI** | 94.51 | 86.11 | 92.70 |
| **RFE(SVC)** | 98.18 | 87.56 | 95.85 |
| **RFE(RandomForest)** | 98.18 | 87.56 | 95.85 |

| | | | |
|---|---|---|---|
| **Proposed algorithm** | **98.19** | **88.11** | **95.99** |

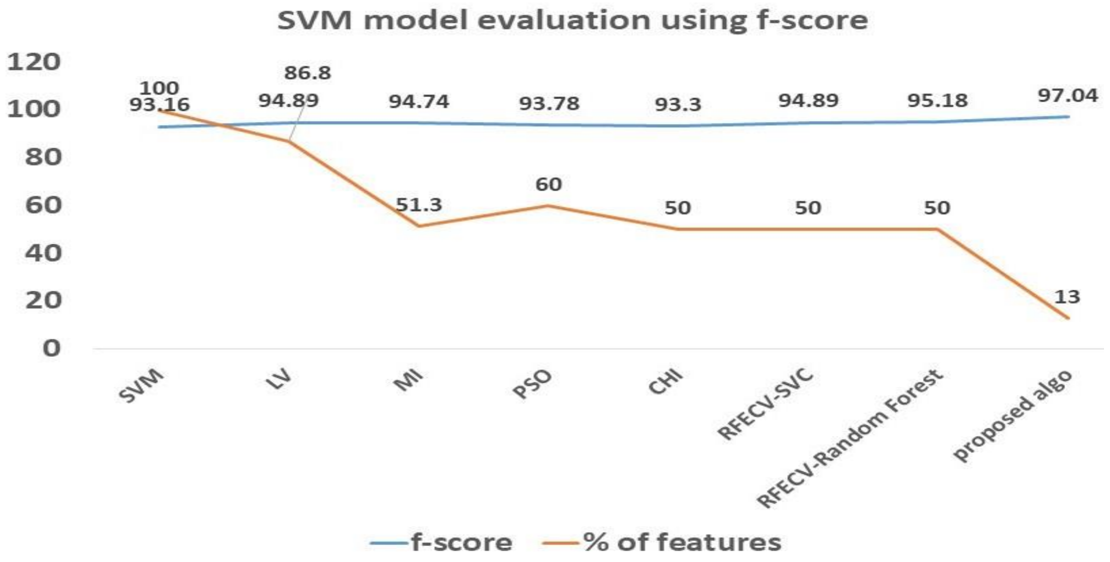
(iii) XGB model with other feature selection algorithm

| Models | Precision | Recall | F-0.5 Score |
|---|---|---|---|
| **Without Feature selection** | 96.20 | 84.44 | 93.59 |
| **With stopwords** | 97.50 | 86.67 | 95.12 |
| **LV** | 98.23 | 90.27 | 96.53 |
| **MI** | 96.93 | 85.40 | 94.38 |
| **PSO** | 96.31 | 84.86 | 93.78 |
| **CHI** | 95.67 | 86.11 | 93.59 |
| **RFE(SVC)** | 98.23 | 90.27 | 96.53 |
| **RFE(RandomForest)** | 98.23 | 90.27 | 96.53 |
| **Proposed algorithm** | **99.41** | **90.81** | **97.56** |

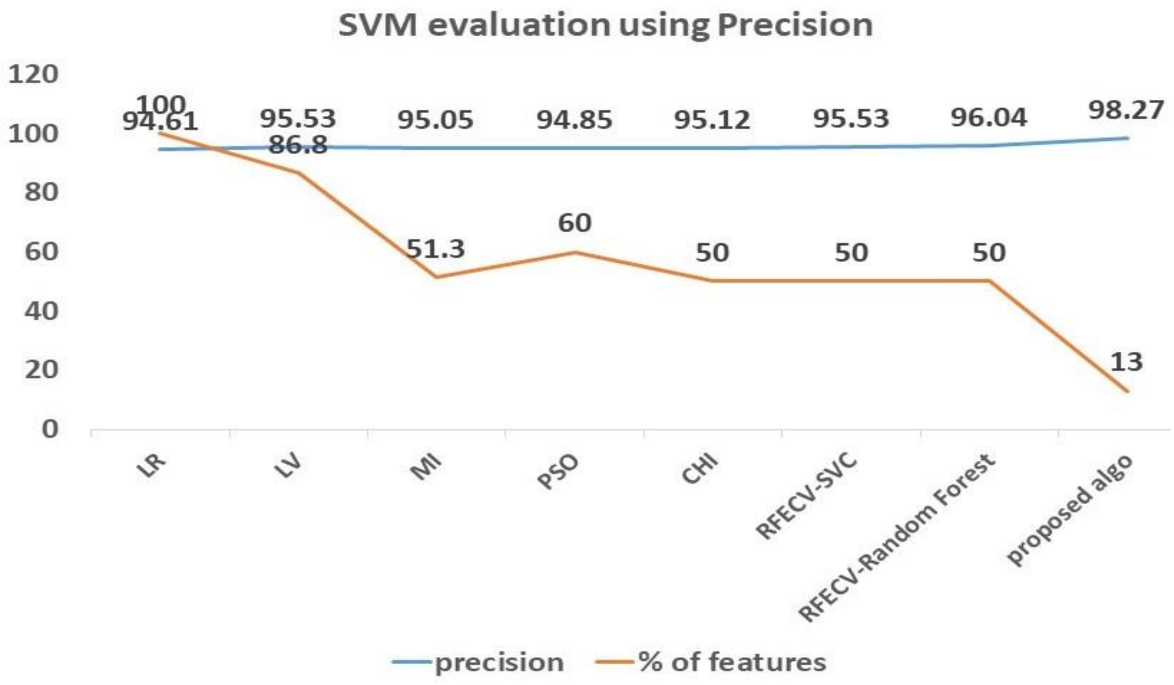
(iv) LSTM model

| Models | Precision | Recall | F-0.5 Score |
|---|---|---|---|
| **Without Feature selection** | 98.79 | 90.27 | 96.88 |

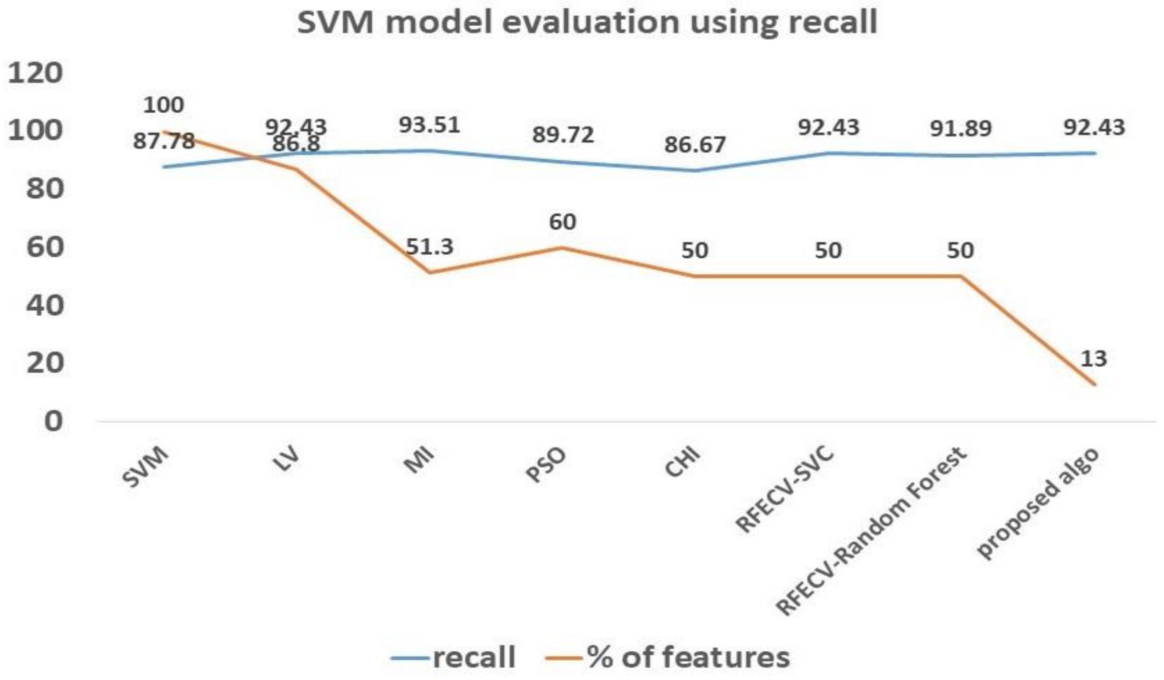
Figure 7.3 gives the bar graph for the comparison study of different feature selection algorithm over Support Vector Machine. Figure 7.4 gives the bar graph for the comparison study of different feature selection algorithm over Logistic Regression. Figure 7.5 gives the bar graph for the comparison study of different feature selection algorithm over XGB model.



(i) F-score against all feature selection algorithm

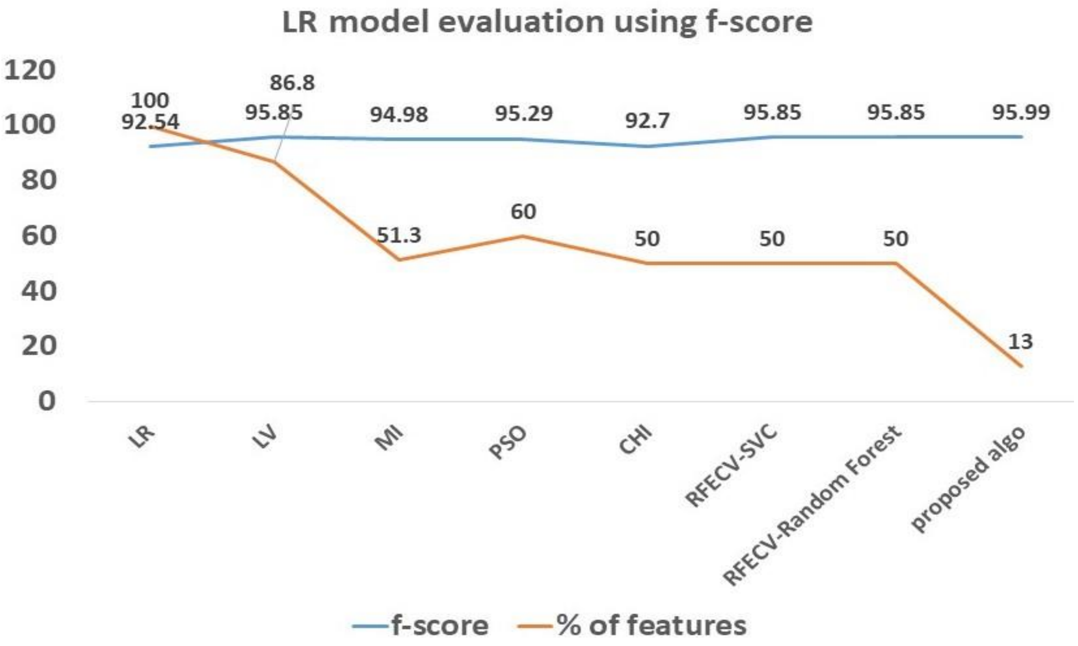


(ii) Precision score against all feature selection algorithm

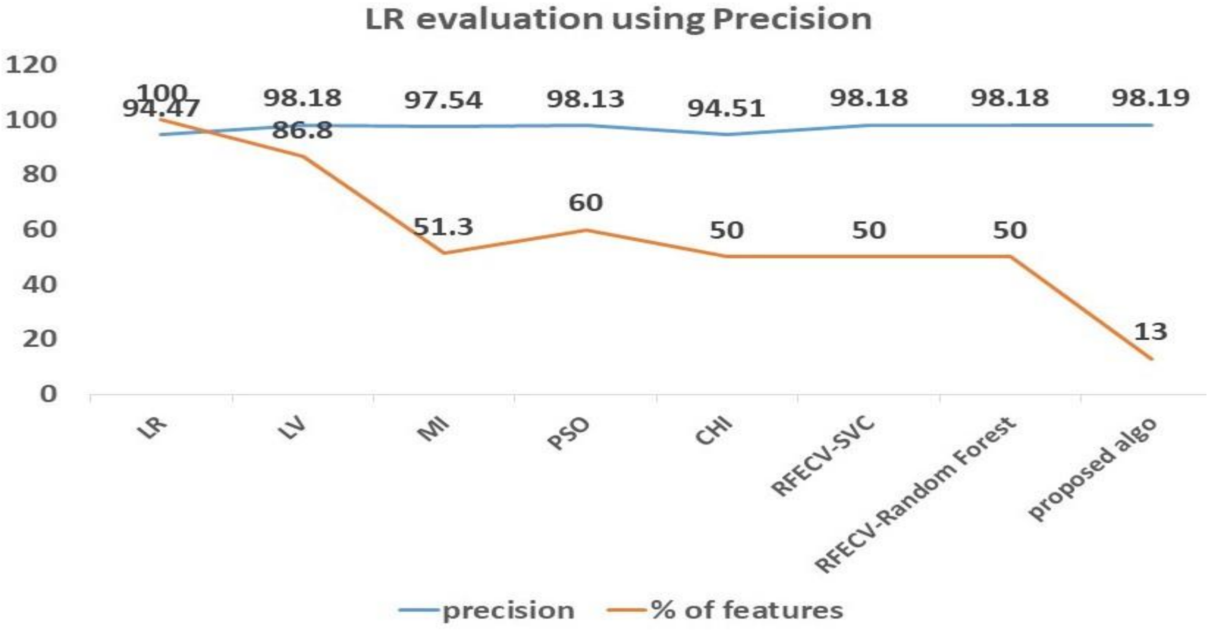


(iii) Recall against all feature selection algorithm

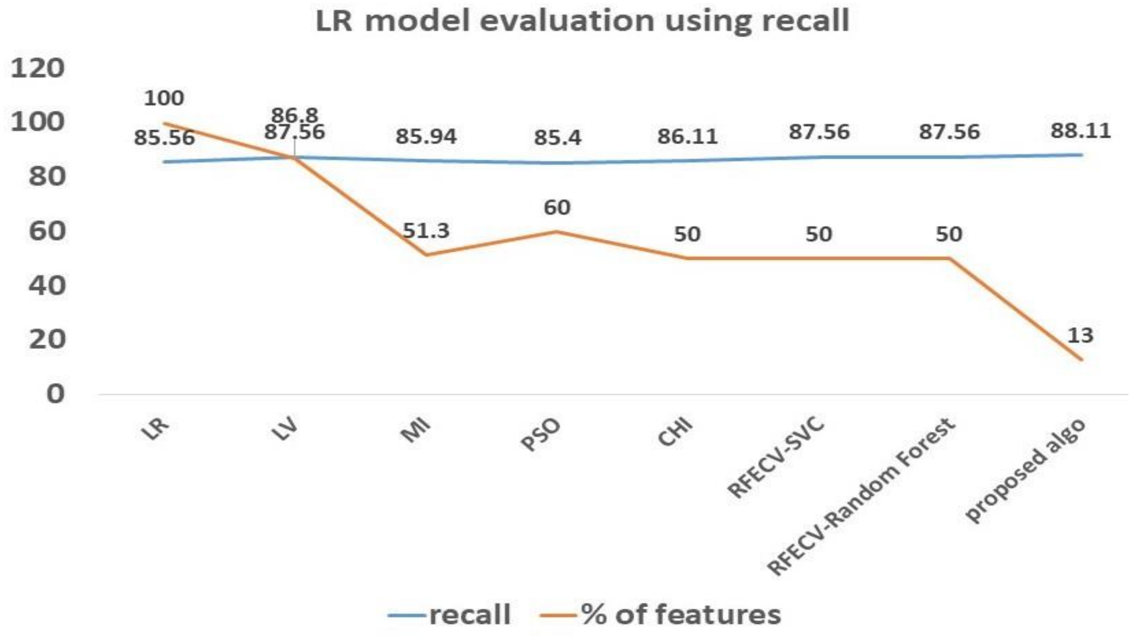
Figure 7.3 Comparative analysis of proposed work with all other feature selection algorithm over SVM model



(i) F-score against all feature selection algorithm

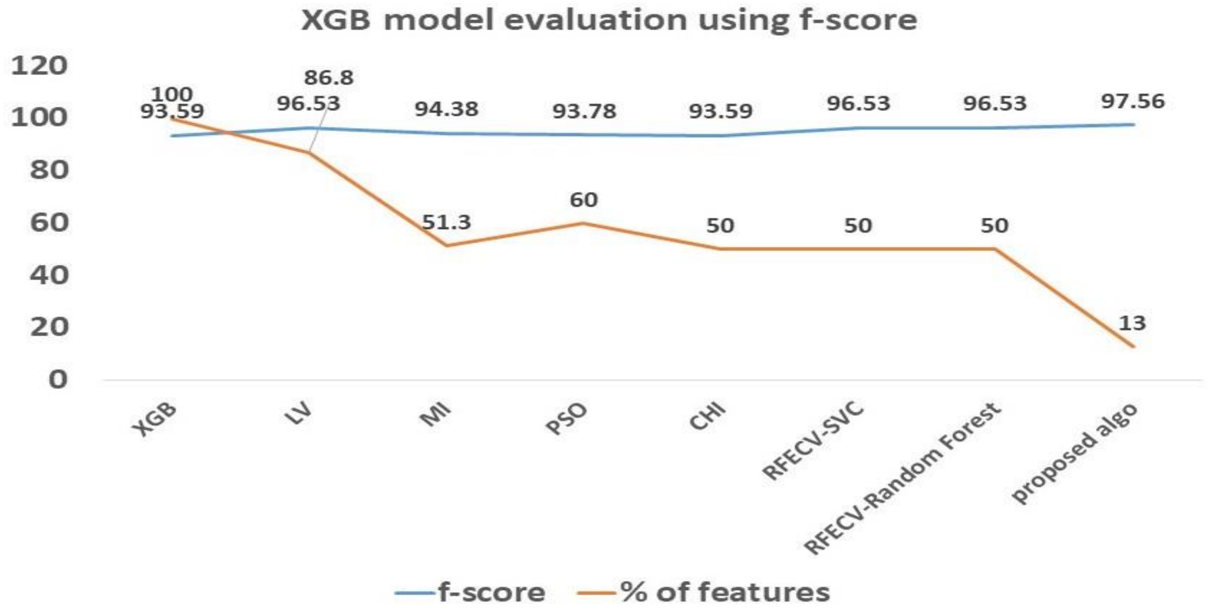


(ii) Precision score against all feature selection algorithm

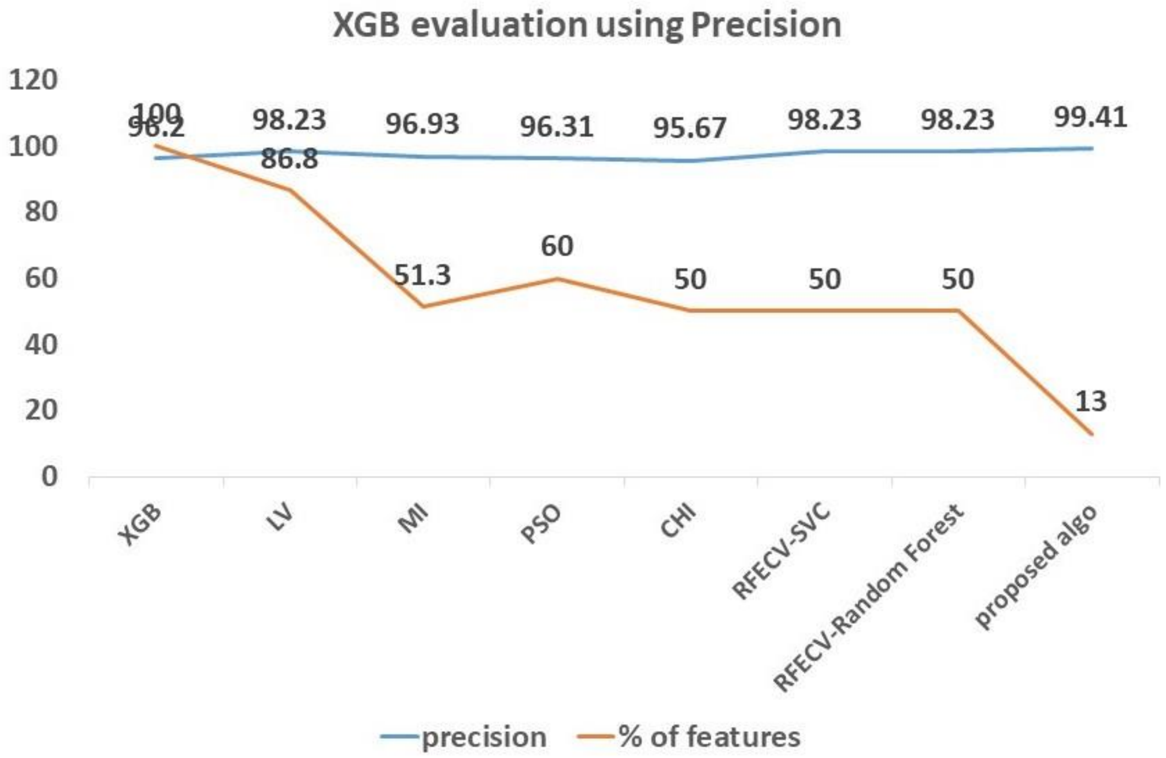


(iii) Recall against all feature selection algorithm

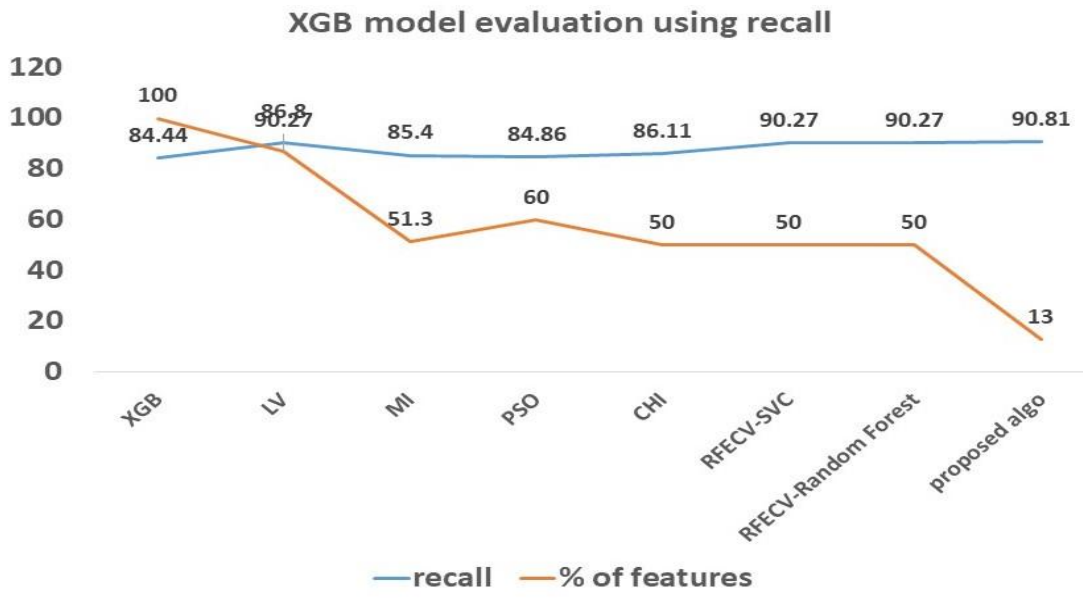
Figure 7.4 Comparative analysis of proposed work with all other feature selection algorithm over LR model



(i) F-score against all feature selection algorithm



(ii) Precision score against all feature selection algorithm



(iii) Recall against all feature selection algorithm

Figure 7.5 Comparative analysis of proposed work with all other feature selection algorithm over XGB model

From the results obtained it can be claimed that proposed work has given more acceptable results than compared with other feature selection algorithms with least number of features. So we can conclude that the two level filtering model proposed in this work is more reliable as it has given the more optimal ten times reduced dataset that can optimally substitute the original dataset. From the above results it could also be concluded that the results achieved by all other algorithms are also comparable with our model which proves the presence of more redundant features than the irrelevant one in the dataset.

So our model is more efficient as it can give more reliable results with such reduced dataset that are alternative to the hard and too complex deep learning models.

# CONCLUSION

As with this invent of new technologies that are being spreading worldwide the large amount of data that is generated has be to classified against the rogue data is generated everyday in the server or network. Though the machine learning models are there that are available to classify the data but the large number of features that are present in the data may lead to misclassification or more computation cost. As there are large number of irrelevant or redundant features that are present in the model this study aims to remove such features. The work proposed in this study is to eliminate these features and produce the most optimal reduced subset that can corresponds the original dataset. The algorithm proposed has given acceptable results over the three chosen machine learning models also the results were able to outperform the deep learning LSTM model. The proposed algorithm with almost 10 times reduced dataset had given quality results that proves the efficacy of the proposed feature selection algorithm.

The proposed algorithm serves as an exemplary of multilevel filtering that could be employed in real world for better classification of such unsolicited messages from the network.

# FUTURE SCOPE

The advancement of technology is bringing new techniques and researches periodically. These are inducing  a large amount of data in the environment that needs some more manageable and strong automated way to throw out such high volume of undesired data. Here we have given a feature selection algorithm based on ensemble learning. The future need some models that could provide more stronger and efficient results as spammers are finding the gateways to break the current security system. Some language independent models and cross domain and platform models are also the prerequisite of the future security models.

# REFERENCES

[1]. Almeida, T. A., Hidalgo, J. M. G., & Yamakami, A. (2011, September). Contributions to the study of SMS spam filtering: new collection and results. In Proceedings of the 11th ACM symposium on Document engineering (pp. 259-262).

[2]. Kalousis, A., Prados, J., & Hilario, M. (2007). Stability of feature selection algorithms: a study on high-dimensional spaces. *Knowledge and information systems*, *12*(1), 95-116.

[3]. Kumar, V. (2014). Sonajharia Minz,“. *Feature Selection: A literature Review”, Smart Computing Review*, *4*(3), 211-229.

[4]. Kohavi, R., & John, G. H. (1997). Wrappers for feature subset selection. *Artificial intelligence*, *97*(1-2), 273-324.

[5]. El Aboudi, N., & Benhlima, L. (2016, September). Review on wrapper feature selection approaches. In *2016 International Conference on Engineering & MIS (ICEMIS)* (pp. 1-5). IEEE.

[6]. Sánchez-Marono, N., Alonso-Betanzos, A., & Tombilla-Sanromán, M. (2007, December). Filter methods for feature selection–a comparative study. In *International Conference on Intelligent Data Engineering and Automated Learning* (pp. 178-187). Springer, Berlin, Heidelberg.

[7]. Porkodi, R. (2014). Comparison of filter based feature selection algorithms: an overview. International journal of Innovative Research in Technology & Science, 2(2), 108-113.

[8]. Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of machine learning research*, *3*(Mar), 1157-1182.

9. Fonti, V., & Belitser, E. (2017). Feature selection using lasso. VU Amsterdam research paper in business analytics, 30, 1-25.

[10]. Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., & Liu, H. (2017). Feature selection: A data perspective. ACM Computing Surveys (CSUR), 50(6), 1-45.

[11]. Yan, K., & Zhang, D. (2015). Feature selection and analysis on correlated gas sensor data with recursive feature elimination. Sensors and Actuators B: Chemical, 212, 353-363.

[12]. Granitto, P. M., Furlanello, C., Biasioli, F., & Gasperi, F. (2006). Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products. Chemometrics and intelligent laboratory systems, 83(2), 83-90.

[13]. Kennedy, J., & Eberhart, R. (1995, November). Particle swarm optimization. In Proceedings of ICNN'95-international conference on neural networks (Vol. 4, pp. 1942-1948). IEEE

[14]. Wang, D., Tan, D., & Liu, L. (2018). Particle swarm optimization algorithm: an overview. Soft Computing, 22(2), 387-408.

[15]. Miranda, L. J. (2018). PySwarms: a research toolkit for Particle Swarm Optimization in Python. Journal of Open Source Software, 3(21), 433.

[16]. Vieira, S. M., Mendonça, L. F., Farinha, G. J., & Sousa, J. M. (2013). Modified binary PSO for feature selection using SVM applied to mortality prediction of septic patients. Applied Soft Computing, 13(8), 3494-3504.

[17]. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735.

[18]. Greff, K., Srivastava, R. K., Koutník, J., Steunebrink, B. R., & Schmidhuber, J. (2016). LSTM: A search space odyssey. IEEE transactions on neural networks and learning systems, 28(10), 2222-2232.

[19]. https://colah.github.io/posts/2015-08-Understanding-LSTMs/

[20]. Soman, K. P., Loganathan, R., & Ajay, V. (2009). Machine learning with SVM and other kernel methods. PHI Learning Pvt. Ltd..

[21]. Auria, L., & Moro, R. A. (2008). Support vector machines (SVM) as a technique for solvency analysis

[22]. Mechelli, A., & Viera, S. (Eds.). (2019). *Machine learning: methods and applications to brain disorders*. Academic Press.

[23]. Menard, S. (2002). Applied logistic regression analysis (Vol. 106). Sage.

[24]. Kleinbaum, D. G., Dietz, K., Gail, M., Klein, M., & Klein, M. (2002). Logistic regression (p. 536). New York: Springer-Verlag.

[25]. Nick, T. G., & Campbell, K. M. (2007). Logistic regression. Topics in biostatistics, 273-301.

[26]. Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794).

[27]. Mussa, D. J., & Jameel, N. G. M. (2019). Relevant SMS spam feature selection using wrapper approach and XGBoost algorithm. Kurdistan Journal of Applied Research, 4(2), 110-120.

[28]. Mustapha, I. B., Hasan, S., Olatunji, S. O., Shamsuddin, S. M., & Kazeem, A. (2020). Effective Email Spam Detection System using Extreme Gradient Boosting. arXiv preprint arXiv:2012.14430.

[29]. Dong, X., Yu, Z., Cao, W., Shi, Y., & Ma, Q. (2020). A survey on ensemble learning. Frontiers of Computer Science, 14(2), 241-258.

[30]. Sagi, O., & Rokach, L. (2018). Ensemble learning: A survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 8(4), e1249.