SPAM MAIL DETECTION USING NATURAL LANGUAGE PROCESSING

B.Ramya
Department of Electrical and Electronics Engineering SR University, Warangal, Telangana, India.
B.Raju
Department of Computer Science and Engineering, SR University, Warangal, Telangana, India.
D.Sindhu Sri
Department of Computer Science and Engineering, SR University, Warangal, Telangana, India.

ABSTRACT:

Email Spam has become a major problem nowadays, with Rapid growth of internet users, Email spams is also increasing. People are using them for illegal and unethical conducts, phishing and fraud. Sending malicious link through spam emails which can harm our system and can also seek in into your system.Spam e-mails can be not only annoying but also dangerous to consumers. Spam e-mail are message randomly sent to multiple addressees by all sorts of groups, but mostly lazy advertisers and criminals who wish to lead you to phishing sites. Spamming is the use of electronic messaging systems like e-mails and other digital delivery systems and broadcast media to send unwanted bulk messages indiscriminately.Implementing spam filtering is extremely important for any organization. Not only does spam filtering help keep garbage out of email inboxes, it helps with the quality of life of business emails because they run smoothly and are only used for their desired purpose.Natural Language Processing (NLP) is a field of computer science and linguistics concerned with the interactions between computers and human (natural) languages we use Neural networks algorithm to detect spam mails.

Keywords:spamdetection,features,phishingsites,unethical,internetusers,lazyadvertisers,algorithms.

1.INTRODUCTION

Email is a means of information transfer from any part of the world that is extremely fast and cost effective and can be used from personal computers, smartphones, and other last-generation electronic gadgets. Despite the increase in usage of other forms of online communication such as instant messaging and social networking, emails have continued to take the lead in business communications and still serves as a requirement for other forms of communications and e-transactions. Emails are used by almost all humans. It is estimated that by the end of 2016, there will be over 2.6 billion email account holders worldwide and it is estimated that nearly half of the world population will be using emails by the end of 2020.

The increase in the popularity and use of emails for transactions has led to a rise in the amount of spam emails globally. Spam emails also called junk emails are unsolicited messages that is non-requested and are almost identical sent to multiple recipients via emails. The sender of spam mails has no prior relationship with the receivers but gathers the addresses from different sources such as phone books and filled forms. Spam messages are fast growing to be one of the most serious threats to users of E-mail messages because it is a major means of sending threats, including viruses, worms and phishing attacks.

Natural Language Processing (NLP) is a field of computer science and linguistics concerned with the interactions between computers and human (natural) languages. Spam e-mails can be not only annoying but also dangerous to consumers. Spam e-mail are message randomly sent to multiple addressees by all sorts of groups, but mostly lazy advertisers and criminals who wish to lead you to phishing sites. Spamming is the use of electronic messaging systems like e-mails and other digital delivery systems and broadcast media to send unwanted bulk messages indiscriminately.

The aim of the paper is to evaluate the performance of Natural Language Processing that are used for grouping emails as spam or not spam.

2.LITERATURE SURVEY

In [1] Unsolicited emails such as phishing and spam emails cost businesses and individuals millions of dollars annually. In this work, the effectiveness of word embedding in classifying spam emails is introduced. Pre-trained transformer model BERT (Bidirectional Encoder Representations from Transformers) is fine-tuned to execute the task of detecting spam emails from non-spam (HAM). BERT uses attention layers to take the context of the text into its perspective. Results are compared to a baseline DNN (deep neural network) model that contains a BiLSTM (bidirectional Long Short Term Memory) layer and two stacked Dense layers. In addition results are compared to a set of classic classifiers k-NN (k-nearest neighbors) and NB (Naive Bayes). Two open-source data sets are used, one to train the model and the other to test the persistence and robustness of the model against unseen data. The proposed approach attained the highest accuracy of 98.67% and 98.66% F1 score.

In [2], Spam or unsolicited emails that are used by spammers can cause huge loss to both the email users and the email server. Therefore, in order to detect spam emails not to enter into our mailbox, a developed email spam classification system is required. This paper proposes two popular machine learning methods, Naïve Bayes Classifier and Support Vector Machine, to classify the emails into spam or ham based on the body or content of the emails. In Naïve Bayes Classifier, independent words are considered as features. Support Vector Machine can be used to represent an email in vector space in which each feature means one dimension. Finally, two methods are compared in terms of precision, recall, F-measure performance metrics with the aim of finding the best method.

In [3], The increase in the use of email in every day transactions for a lot of businesses or general communication due to its cost effectiveness and efficiency has made emails vulnerable to attacks including spamming. Spam emails also called junk emails are unsolicited messages that are almost identical and sent to multiple recipients randomly. In this study, a performance analysis is done on some classification algorithms including: Bayesian Logistic Regression, Hidden Naï ve Bayes, Radial Basis Function

(RBF) Network, Voted Perceptron, Lazy Bayesian Rule, Logit Boost, Rotation Forest, NNge, Logistic Model

Tree, REP Tree, Naïve Bayes, Multilayer Perceptron, Random Tree and J48Rotation Forest is found to be the classifier that gives the best accuracy of 94.2%.

In [4], Unsolicited e-mail also known as Spam has become a huge concern for each e-mail user. In recent times, it is very difficult to filter spam emails as these emails are produced or created or written in a very special manner so that anti-spam filters cannot detect such emails. This paper compares and reviews performance metrics of certain categories of supervised machine learning techniques such as SVM (Support Vector Machine), Random Forest, Decision Tree, CNN, (Convolutional Neural Network), KNN (K Nearest Neighbor), MLP (Multi-Layer Perceptron), Ad boost (Adaptive Boosting) Naïve Bayes algorithm to predict or classify into spam emails. The objective of this study is to consider the details or content of the emails, learn a finite dataset available and to develop a classification model that will be able to predict or classify whether an e-mail is spam or not.

In [5], In this work, the effectiveness of word embedding in classifying spam emails is introduced. Pre-trained transformer model BERT (Bidirectional Encoder Representations from Transformers) is fine-tuned to execute the task of detecting spam emails from non-spam (HAM). BERT uses attention layers to take the context of the text into its perspective. Results are compared to a set of classic classifiers k-NN Two open-source data sets are used, one to train the model and the other to test the persistence and robustness of the model against unseen data. The proposed approach attained the highest accuracy of 98.67% and 98.66% F1 score.

In [6], This paper aims to compare different classifying techniques on different datasets collected from previous research works, and evaluate them on the basis of their accuracy, recall, and precision. The comparison has been performed between traditional machine learning techniques. Most of the time such emails are commercial. But many times, such emails may contain some phishing links that have malware. This arises the need for proposing prudent mechanism to detect or identify such spam emails so that time and memory space of the system can be saved up to a great extent. In this paper, we presented the NLP mechanism which can filter spam and non-spam emails and also

categorize into different spam mails. Our proposed algorithm generates dictionary and features and trains them through machine learning for effective results.

In [7], In this research, we'll look at a strategy that employs natural language processing (NLP) to detect Spam and Ham News using the Spam Email Dataset. We have used Dense classifier Sequential Neural Network , LSTM and Bi- LSTM and compared accuracies and results. The dataset's efficacy is determined using metrics such as recall, accuracy, and F1-score.

In [8], With the increase usage of emails, this study focuses on using automated ways to detect spam emails written in Urdu. study uses various machine learning and deep learning algorithms to detect. deep learning algorithm (LSTM) is a stronger method for detecting Urdu spam emails, with high accuracy of 98.4%.SVM and Naive Bayes. CNN and LSTM are used.

In [9], authors focused on methods to effectively refine SMS spam. Various machine learning based classifiers like the Naïve Bayes, Gradient Boost Logistic Regression, SGD classifier and Deep learning-based models like CNN and LSTM, were also tested. As per their results, CNN model with the regularisation parameter on randomly generated tenfold cross validation data worked best in terms of filtering legitimate text messages, having an accuracy of 99.44%

In [10], compared various machine learning algorithms such as SVM (Support Vector Machine), Random Forest, Decision Tree, CNN (Convolutional Neural Network), KNN(K Nearest Neighbor), MLP(Multi-Layer Perceptron), Adaboost (Adaptive Boosting) ,Naïve Bayes algorithm to predict or classify into spam emails. There are various SPAM datasets such as SPAM ARCHIVE, SPAMBASE, LINGSPAM etc to perform this experiment. We used SPAMBASE dataset to evaluate performance of above algorithms in terms of Accuracy, Precision

3.PROBLEM DEFINATION

This is a study of the problem of spam mail detection using NLP.

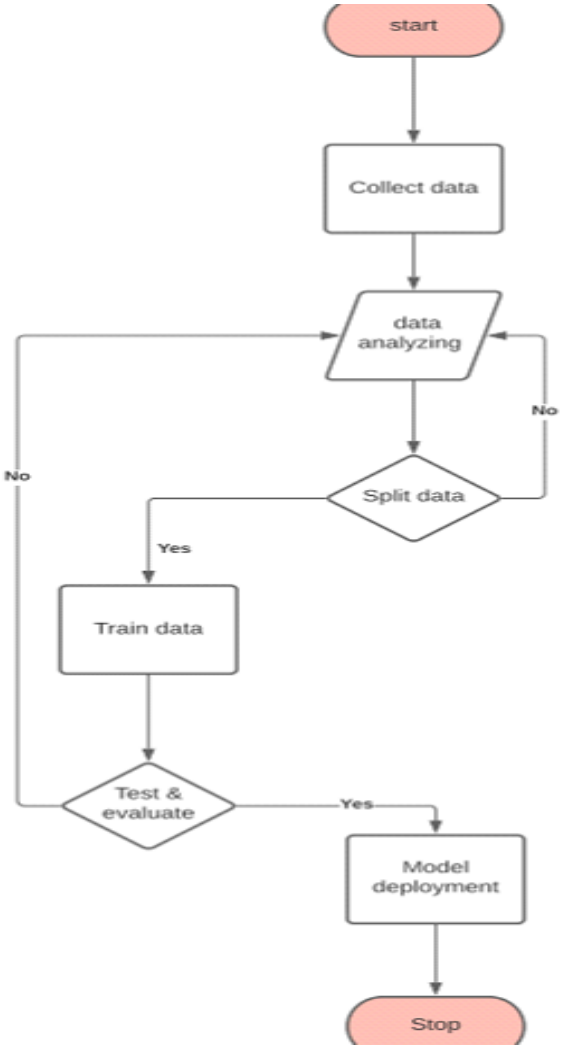

Figure 2. Processing Steps for spam mail detectio

4. DATASET AND ATTRIBUTES

	Category	Message
2	ham	Go until jurong point, crazy.. Available only in bugis n great world la e buffet... Cine there got amore wat...
3	ham	Ok lar... Joking wif u oni...
4	spam	Free entry in 2 a wkly comp to win FA Cup final tkts 21st May 2005. Text FA to 87121 to receive entry question(std txt r
5	ham	U dun say so early hor... U c already then say...
6	ham	Nah I don't think he goes to usf, he lives around here though
7	spam	FreeMsg Hey there darling it's been 3 week's now and no word back! I'd like some fun you up for it still? Tb ok! XxX st
8	ham	Even my brother is not like to speak with me. They treat me like aids patent.
9	ham	As per your request 'Melle Melle (Oru Minnaminunginte Nurungu Vettam)' has been set as your callertune for all Call
10	spam	WINNER!! As a valued network customer you have been selected to receivea £900 prize reward! To claim call 0906170
11	spam	Had your mobile 11 months or more? U R entitled to Update to the latest colour mobiles with camera for Free! Call Th
12	ham	I'm gonna be home soon and i don't want to talk about this stuff anymore tonight, k? I've cried enough today.
13	spam	SIX chances to win CASH! From 100 to 20,000 pounds txt> CSH11 and send to 87575. Cost 150p/day, 6days, 16+ TsandCs
14	spam	URGENT! You have won a 1 week FREE membership in our £100,000 Prize Jackpot! Txt the word: CLAIM to No: 81010 T&
15	ham	I've been searching for the right words to thank you for this breather. I promise i wont take your help for granted and
16	ham	I HAVE A DATE ON SUNDAY WITH WILL!!
17	spam	XXXMobileMovieClub: To use your credit, click the WAP link in the next txt message or click here>> http://wap. xxxm
18	ham	Oh k...i'm watching here:)
19	ham	Eh u remember how 2 spell his name... Yes i did. He v naughty make until i v wet.
20	ham	Fine if that□s the way u feel. That□s the way its gota b
21	spam	England v Macedonia - dont miss the goals/team news. Txt ur national team to 87077 eg ENGLAND to 87077 Try:WALES
22	ham	Is that seriously how you spell his name?
23	ham	I'm going to try for 2 months ha ha only joking
24	ham	So ü pay first lar... Then when is da stock comin...
25	ham	Aft i finish my lunch then i go str down lor. Ard 3 smth lor. U finish ur lunch already?
26	ham	Ffffffffff. Alright no way I can meet up with you sooner?

Figure 2 Visualizing attributes of the dataset

5. DATA PRE-PROCESSING

Dataset is a collection of data or related information that is composed for separate elements. A collection of dataset for e-mail spam contains spam and non-spam messages. Dataset (rows X columns) 5573 X 2.

Category	String
Message	String Large text data

Table 3. Attributes Validation

Dataset consists of spam and non-spam messages and 5573 rows and 2 columns one column describes about the mail received by the user and another column contains whether the received mail is spam or not.

Stemming:

Stemming is the process of reducing a word to its word stem that affixes to suffixes and prefixes or to the roots of words known as a lemma. Stemming is important in natural language understanding (NLU) and natural language processing (NLP). Stemming is a part of linguistic studies in morphology and artificial intelligence (AI) information retrieval and extraction. Stemming and AI knowledge extract meaningful information from vast sources like big data or the Internet since additional forms of a word related to a subject may need to be searched to get the best results Recognizing, searching and retrieving more forms of words returns more results. When a form of a word is recognized it can make it possible to return search results that otherwise might have been missed. That additional information retrieved is why stemming is integral to search queries and information retrieval.

Removing stop words:

The words which are generally filtered out before processing a natural language are called stop words. These are actually the most common words in any language (like articles,

prepositions, pronouns, conjunctions, etc) and does not add much information to the text. Examples of a few stop words in English are “the”, “a”, “an”, “so”, “what”. Stop words are available in abundance in any human language. By removing these words, we remove the low-level information from our text in order to give more focus to the important information. In order words, we can say that the removal of such words does not show any negative consequences on the model we train for our task. Removal of stop words definitely reduces the dataset size and thus reduces the training time due to the fewer number of tokens involved in the training.

NLP is one of the most researched areas today and there have been many revolutionary developments in this field. NLP relies on advanced computational skills and developers across the world have created many different tools to handle human language. Out of so many libraries out there, a few are quite popular and help a lot in performing many different NLP tasks.

Tokenization:

Tokenization is the first step in any NLP pipeline. It has an important effect on the rest of your pipeline. A tokenizer breaks unstructured data and natural language text into chunks of information that can be considered as discrete elements. The token occurrences in a document can be used directly as a vector representing that document. This immediately turns an unstructured string (text document) into a numerical data structure suitable for machine learning. They can also be used directly by a computer to trigger useful actions and responses. Or they might be used in a machine learning pipeline as features that trigger more complex decisions or behavior.

punctuation removal:

The punctuation removal process will help to treat each text equally. For example, the word data and data! are treated equally after the process of removal of punctuations. We need to take care of the text while removing the punctuation because the contraction words will not have any meaning after the punctuation removal process. Such as ‘don’t’ will convert to ‘dont’ or ‘don t’ depending upon what you set in the parameter.We also need to be extra careful while choosing the list of punctuations that we want to exclude from the data

depending upon the use cases. As string.punctuation in python contains these symbols !"#$%&\'()*+,-./:;?@[\\]^_{|}~`

6. ALGORITHMS

This section talks about the algorithms used for the project. We used LSTM(Long Short Term Memory) and Word to Vector.

6.1 LSTM(LONG SHORT TERM MEMORY)

deep learning model based LSTM(Long short term memory) method for spam detection in email.

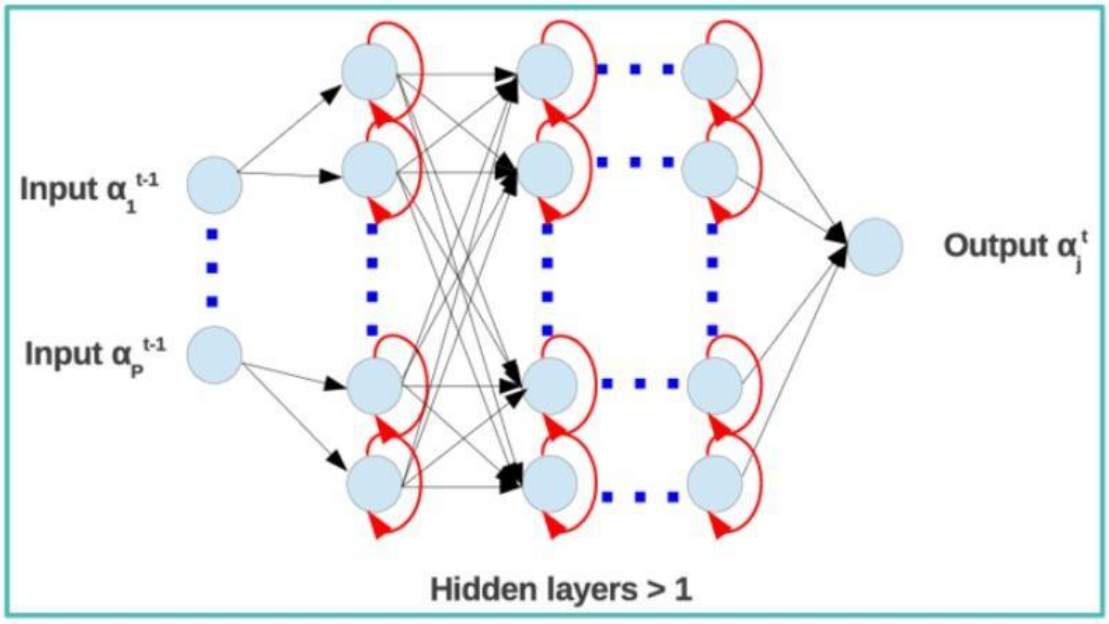

Fig 4:LSTM Model

LSTMs have three types of gates they are:

Input gates, forget gates, and output gates which controls the flow of information. The hidden layer output of LSTM includes the hidden state and the memory cell. Only the hidden state is passed into the output layer. The memory cell is entirely internal.

This challenge to address long-term information preservation and short-term input skipping in latent variable models has existed for such a long time. One of the earliest approaches to

address this was the long short-term memory (LSTM) . It shares many of the properties of the GRU. Interestingly, LSTMs have a slightly more complex design than GRUs but predates GRUs by almost two decades.

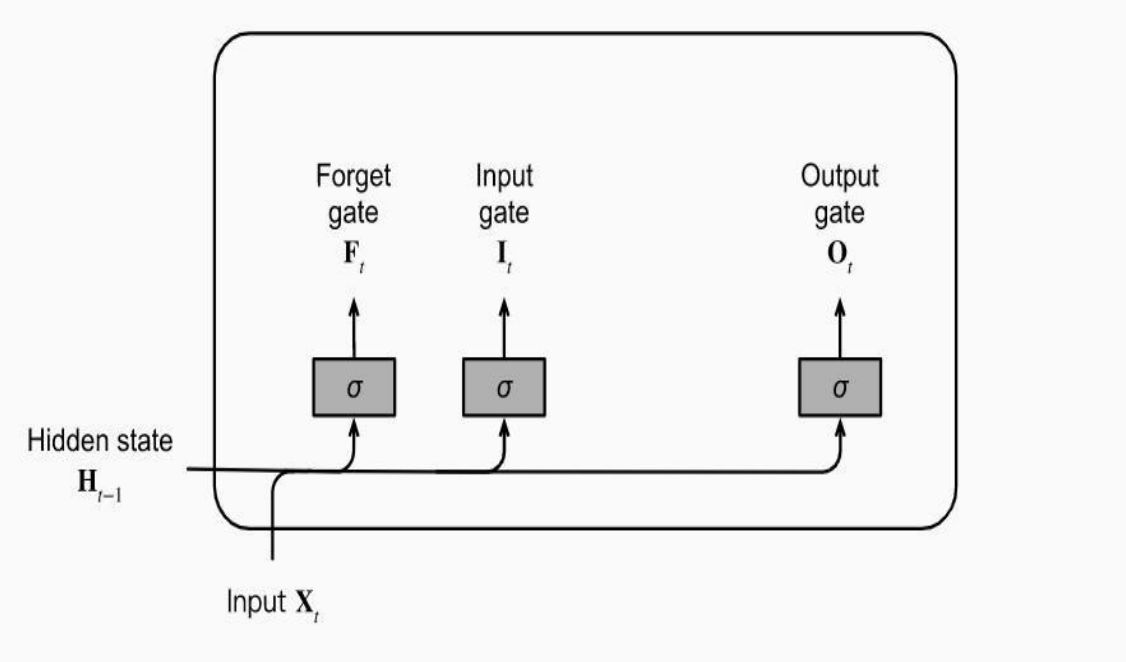

Fig.5: gated memory cell in lstm

6.2 WORD TO VECTOR

Word2Vec model is used for Word representations in Vector Space which is founded by Tomas Mikolov and a group of the research teams from Google in 2013. It is a neural network model that attempts to explain the word embeddings based on a text corpus.

These models work using context. This implies that to learn the embedding, it looks at nearby words; if a group of words is always found close to the same words, they will end up having similar embeddings. To label how words are similar or close to each other, we first fix the window size, which determines which nearby words we want to pick.

The General Flow of the Algorithm

Step-1: Initially, we will assign a vector of random numbers to each word in the corpus.

Step-2: Then, we will iterate through each word of the document and grab the vectors of the nearest n-words on either side of our target word, and concatenate all these vectors, and then

forward propagate these concatenated vectors through a linear layer + softmax function, and try to predict what our target word was.

Step-3: In this step, we will compute the error between our estimate and the actual target word and then backpropagated the error and then modifies not only the weights of the linear layer but also the vectors or embeddings of our neighbor's words.

Step-4: Finally, we will extract the weights from the hidden layer and by using these weights encode the meaning of words in the vocabulary.

Word2Vec model is not a single algorithm but is composed of the following two pre-processing modules or techniques:

Continuous Bag of Words (CBOW)

Skip-Gram.

Both of the mentioned models are basically shallow neural networks that map word(s) to the target variable which is also a word(s). These techniques learn the weights that act as word vector representations. Both these techniques can be used to implementing word embedding using word2vec.

Why Word2Vec technique is created?

As we know that most of the NLP systems treat words as atomic units. In existing systems with the same purpose as that of word2vec, there is a disadvantage that there is no notion of similarity between words. Also, those system works for small, simpler data and outperforms on because of only a few billions of data or less.

So, In order to train the system with a larger dataset with complex models, these techniques use a neural network architecture to train complex data models and outperform huge datasets with billions of words and with vocabulary having millions of words.

It helps to measure the quality of the resulting vector representations and works with similar words that tend to close with words that can have multiple degrees of similarity.

Syntactic Regularities: These regularities refer to grammatical sentence correction.

Semantic Regularities: These regularities refer to the meaning of the vocabulary symbols arranged in that structure.

The proposed technique was found that the similarity of word representations goes beyond syntactic regularities and works surprisingly well for algebraic operations of word vectors.

Continuous Bag of Words (CBOW)

The aim of the CBOW model is to predict a target word in its neighborhood, using all words. To predict the target word, this model uses the sum of the background vectors. For

this, we use the pre-defined window size surrounding the target word to define the neighboring terms that are taken into account.

Advantages of CBOW:

1. Generally, it is supposed to perform superior to deterministic methods due to its probabilistic nature.
2. It does not need to have huge RAM requirements. So, it is low on memory.

Disadvantages of CBOW:

1. CBOW takes the average of the context of a word. For Example, consider the word apple that can be both a fruit and a company but CBOW takes an average of both the contexts and places it in between a cluster for fruits and companies.
2. If we want to train a CBOW model from scratch, then it can take forever if we not properly optimized it.

Skip-Gram

1. Given a word, the Skip-gram model predicts the context.
2. Skip–gram follows the same topology as CBOW. It just flips CBOW's architecture on its head. Therefore, the skip-gram model is the exact opposite of the CBOW model.
3. In this case, the target word is given as the input, the hidden layer remains the same, and the output layer of the neural network is replicated multiple times to accommodate the chosen number of context words.

Advantages of Skip-Gram Model

1. The Skip-gram model can capture two semantics for a single word. i.e two vector representations for the word Apple. One for the company and the other for the fruit.
2. Generally, Skip-gram with negative sub-sampling performs well then every other method.

7. RESULTS

```
score = model.evaluate(x_test, yval, batch_size=32)
print()
print("ACCURACY:",score[1])
print("LOSS:",score[0])
```

```
35/35 [==============================] - 4s 101ms/step - loss: 0.0727 - accuracy: 0.9812

ACCURACY: 0.9811659455299377
LOSS: 0.07265845686197281
```

Figure 8. Result of various models with the proposed model

The neural network deep learning algorithms that we used is LSTM(long-short term memory). This algorithms worked well on spam mail detection. We got 98% accuracy foulds. LSTM has four layers

1)LSTM layer-1

2)LSTM layer-2

3)drop out layer

4)dense layer.

8. CONCLUSION:

The previous or existing spam mail detection systems used traditional text based machine learning models. The results highly rely on the crafted extracted features. The performances are unstable when detecting spam mails.

So, we propose a deep learning model based LSTM(Long short term memory) method for spam mail detection. The neural network deep learning algorithms that we used is LSTM(long-short term memory). This algorithms worked well on spam mail detection system. We got 98% accuracy.

During computation of long text mails which are far away, it is impossible to store which causes vanishing of gradient. In order to maintain we use LSTM.

10.REFERENCES

1. Yaseen, Qussai. "Spam email detection using deep learning techniques." Procedia Computer Science 184 (2021): 853-858.
2. T. M. Ma, K. YAMAMORI and A. Thida, "A Comparative Approach to Naïve Bayes Classifier and Support Vector Machine for Email Spam Classification," 2020 IEEE 9th Global Conference on Consumer Electronics (GCCE), 2020, pp. 324-326, doi: 10.1109/GCCE50665.2020.9291921.
3. Shafi'i, Muhammad Abdulhamid, et al. "Comparative analysis of classification algorithms for email spam detection." (2018).
4. https://www.researchgate.net/publication/352477326_Prediction_of_Spam_Email_using_Machine_Learning_Classification_Algorithm.
5. https://www.researchgate.net/publication/351678576_Spam_Email_Detection_Using_Deep_Learning_Techniqueshttps://iarjset.com/papers/email-spam-detection-using-machine-learning-techniques/
6. https://www.researchgate.net/publication/357154618_Machine_Learning-Based_Detection_of_Spam_Emails
7. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4145123
8. P. K. Roy, J. P. Singh, and S. Banerjee, "Deep learning to filter SMS Spam," Futur. Gener. Comput. Syst., vol. 102, pp. 524–533, Jan. 2020, doi: 10.1016/j.future.2019.09.001.
9. P. K. Roy, J. P. Singh, and S. Banerjee, "Deep learning to filter SMS Spam," Futur. Gener. Comput. Syst., vol. 102, pp. 524–533, Jan. 2020, doi: 10.1016/j.future.2019.09.001.
10. EmmanuelGbengaDadaa, Joseph StephenBassia, HarunaChiromab, Shafi'i MuhammadAbdulhamidc, Adebayo OlusolaAdetunmbid, Opeyemi EmmanuelAjibuwae: "Machine learning for email spam filtering: review,approaches and open research problems",Heliyon, Volume 5, Issue 6, June 2019
11. https://link.springer.com/article/10.1007/s41237-021-00142-y