Spam-Detection with Comparative Analysis and Spamming Words Extractions

This paper was downloaded from TechRxiv (<https://www.techrxiv.org>).

LICENSE

CC BY 4.0

SUBMISSION DATE / POSTED DATE

19-10-2021 / 21-10-2021

CITATION

Islam, Md Khairul; Amin, Md Al; Islam, Md Rakibul; Mahbub, Md Nosin Ibna; Showrov, Md Imran Hossain; Kaushal, Chetna (2021): Spam-Detection with Comparative Analysis and Spamming Words Extractions. TechRxiv. Preprint. <https://doi.org/10.36227/techrxiv.16832320.v1>

DOI

[10.36227/techrxiv.16832320.v1](https://doi.org/10.36227/techrxiv.16832320.v1)

Spam-Detection with Comparative Analysis and Spamming Words Extractions

Md Khairul Islam
*Information and Communication
Technology, Islamic University
Kushtia, Bangladesh
mdkito51@gmail.com*

Md Al Amin
*Computer Science and Engineering
Prime University
Dhaka, Bangladesh
ahmedalamin2357@gmail.com*

Md Rakibul Islam
*Information and Communication
Technology, Islamic University
Kushtia, Bangladesh
mrpislam@gmail.com*

Md Nosin Ibna Mahbub
*Information and Communication
Technology, Islamic University
Kushtia, Bangladesh
mahbub.nosin@gmail.com*

Md Imran Hossain Showrov
*Institute of Computer Science
Bangladesh Atomic Energy Commission
Dhaka, Bangladesh
showrov@baec.gov.bd*

Chetna Kaushal
*Chitkara University Institute of Engineering
and Technology, Chitkara University
Punjab, India.
chetnakaushal3558@gmail.com*

Abstract—Communication through email plays an essential part especially in every sector of our day-to-day life. Considering its significance, it is important to filter spam emails from emails. Spam email, also known as junk email, is unwanted messages that are sent by the electronic medium in large quantities. Most of the spam emails are commercial in nature that is not only irritating but also harmful due to malicious scams or malware-hosting sites or use viruses attached to the message. In this paper, we identify spam emails and expose how spam emails can be distinguished from legitimate/normal emails. We deployed four machine learning models and two deep learning models over the datasets including the combined dataset. Besides, we also try to find the important keywords that are found repeatedly from spam emails repository. This type of knowledge will enable us to detect spam emails for our personnel and community security purpose.

Index Terms—Spam mail, legitimate, NLP, Word Clouds, Machine Learning, Deep Learning

I. INTRODUCTION

E-mail refers to the sharing of data among persons that use digital networking equipment. In 1971, the very first e-mail is invented and used by Ray Tomlinson. He developed the first system capable of sending mail throughout the ARPANET among users on different hosts, utilizing @ sign to link the mail address at the destination space. And the technology was known as email in the early 1980s [1, 2, 3].

Emails have become an easy technique to share data, ideas, and commonly written correspondence around the world. And it is a way of communicating via electronic devices among individuals. More specifically, e-mail is a file that may contain text, files, photographs, web addresses, or other items that are transmitted over a network to a specific person or a group of people at the same time. For emailing to individuals or groups no additional fees need not be paid. In 2019, the total global subscribers of the e-mail were 3.9 billion, it is expected to grow over 4.48 billion in 2024. In 2018, about 281 billion e-mails globally were being sent and received every

day. Nowadays, E-mail is a convenient and available service that is sensitive as well. It is also available from anywhere, to facilitate bulk message transmission, to enable instant access to information including any types of files. So, email has become one of the world's most common communication system [4, 5].

However, as technology advances, the innovation of email has been hurt by an outfit called email spamming. Email spamming can be done in the form of unintentional, unsolicited electronic contact. Web-users also use e-mails for website and email sign-ups and brace themselves for the potential flood of malware and advertising updates. While most unwanted e-mails are disturbing and essentially innocuous, but users have to be concerned about malicious e-mails that have been sent to hack one's online identity and computer data. Since the early 1990s, email spam has gradually grown and is expected to account for about 90 percent of the worldwide mail transactions by 2014 [6, 7]. There is no need for spam emails that are used to drain space, network connectivity, and consume the time of receivers. Spam emails reported 53.95 percent of e-mail traffic in March 2020. Spam emails usually aim to gather confidential customer/personal details information so that the scammers can target and harass the receiver. So it is important to find out and avoid the junk/spam email from inbox [8, 9].

In this paper, we identified spam emails from baskets of emails using Machine Learning and Deep Learning approaches. We have carried out some unique steps and operations to do so. Such as, we used four different datasets in our paper. Those datasets are Trec spam Dataset-2007, Enron email dataset, PU dataset, Lingspam dataset. And we have used both Machine Learning and Deep Learning Models in our work. Which has given us pretty much nice accuracy. We created word clouds to show spam words that are liable for spamming emails for each dataset. First, we have used the feature engineering technique to find the most important features from the datasets we have collected. Then the data has

been preprocessed for better accuracy of the models. After data preprocessing different techniques are used to convert the preprocessed data into a numerical vector. The datasets are also split as train, test, and cross-validation. We have used several Machine Learning Models such as Logistic Regression, xgboost, Support Vector Machine, Random Forest. And we also used Deep Learning Models such as Word Embedding, LSTM.

The organization of the rest of the paper is as follows. A brief outline of the related literature is presented in section II. A precise explanation of the proposed methods is given in section III. Also, a brief description of the techniques used in the study and dataset are given here. Experimental setup and results are presented in section IV. Finally, the paper finishes with the concluding remarks in section V.

II. RELATED WORKS

There are several study and research works have been done for classifying the spam mail by using various datasets with machine learning and deep learning approach. Such as, for TREC07 dataset a study has been done by using Naïve Bayes and Neural Network (MLP) where limitations was low performance. Sharma, Prajapat and Aslam [10] used the classification algorithm of keywords to implement multilayer perceptron neural network (MLP) and naïve Bayes models. Utilizing statistical analysis on communications, they often calculate their findings as either junk or ham upon this TREC07 dataset. A significant downside of the MLP model is that, relative to NB, the training is sluggish as it costs a lot to develop. Another study has done using Lingspam dataset and in this study only accuracy of the method was used in assessing its performance. Using a local outlier factor (LOF) because as performance feature for the predictor distribution for classification problems on the Ling spam dataset, Palanisamy, Kumaresan and Varalakshmi [11] implemented the hybrid mixed negative selection algorithm (NSA) and PSO. As variables such as accuracy, memory, calculation time, and false positives were not included in measuring the system's performance, the success of their method can not be measured. In determining its efficiency, just the classification precision of the system was used. By using better optimization techniques, the proposed work can be further improved to increase its performance, as the accuracy of the classification is still very poor. PU and Enron Spam datasets are used in a study where the limitation is Time consuming training. Akshita [12] introduced the Deep Learning methodology to spam classification content - based recommendation. On PU1, PU2, PU3, PUA and Enron spam datasets, the writer utilized DL4J network model.

There is another research has been done using both PU1, Ling-spam. Zhou[13] has performed the research work but the limitation is interpretation and computation required for higher and lower spam thresholds. To minimize the chances of misclassification, in updated Naïve Bayesian mail junk filter with cost-effective three significant method is used. Another research has been conducted using TREC, Spam Assassin

datasets. Zhong [14] has done has the research where the limitation is Word obfuscation affects the classification accuracy. C. Hua Li [15] has done a research based on Ling-spam, PU1 and PU3 datasets. However the limitation of the work is training takes more time. Sunita [16] has been done a study using both trec and lingspam dataset using machine learning and deep learning. In the research CNN and different machine learning models are implemented. Sanjiban and Abhishek [17] has done a research with machine learning and deep learning. They used SVM and ANN models for the classification. Ammara and hikmat [18] has done the study using several machine learning and deep learning models. SVM, AdaBoost, MLP, DNN, RF models are used in the research. Shikhar Seth [19] has completed a research on Enron Spam Dataset using deep learning technique. Ankit Narendrakumar Soni[20] also has done research on enron dataset using deep learning technique. LSTM and CNN models are used in the study.

In a recent approach [21], researcher used SVM for classifying spam emails from legitimate emails. They have used Laplace feature map algorithms for extracting important features. Although three different datasets used for this study, but their accuracy wasn't up to the mark. In an another study [22] junk emails filtered using different machine and deep learning approaches. However, word embedding with deep learning performs over any other algorithms. unsupervised learning is also employed for spam filtering. To my surprise, it performs admirably in terms of accuracy [23].

III. METHODOLOGY

This section describes different steps illustrated in figure 1.

A. Data Acquisition and Preprocessing

In this paper, four different datasets have been used both in Machine learning and Deep learning. The datasets are Trec spam Dataset-2007, Enron email dataset, PU123ACorpora dataset, Ling-spam dataset. We also create a basket dataset. And its a combination of all the four datasets we have selected for our case. Dataset then preprocessed in order to fit into algorithms. Dataset description is given in table I.

TABLE I: Dataset description

Dataset	Total emails	Spam	legitimate
Trec spam-2007 [24]	75419	50199	25220
Enron email [25]	33716	17171	16545
PU123ACorpora [26]	7101	3020	4081
Ling-Spam [27]	11572	1924	9648
Basket dataset	127808	72314	55494

B. Feature Extraction and Selection

From trec 2007 dataset, we have extracted these features: Return-Path, Message-ID, Body, Date, From, MIME-Version, Content-Type. For Enron and Ling - spam datasets, we have extracted two features: Subject and Body. The Bag of Words

Fig. 1: A schematic approach of proposed model

technique was already used for simplifying the representation of the PU123ACorpora dataset. No feature extraction technique is used for deep learning models.

C. Summarization and Context Learning

In summarization and context learning, we have introduced two approaches: spam-legitimate email summarization and keyword finding. In the summarization approach, we have identified a summary of all spam emails and the most significant sentences that turned the email into unwanted one. Keyword findings are meant to classify the most important keywords in the review text to visualize the corresponding datasets contextual sense. We have also used wordcloud for keywords for context learning techniques. Because it is possible to highlight important text-based data points using a word cloud. Word clouds are commonly used to examine data of social networks and other resource [28, 29].

D. Classifier Models

In order to identify the spam emails, We used the classifiers. A binary classification algorithm applied to estimate the likelihood of a target variable is logistic regression. The existence of the objective or dependent variable is dichotomous, meaning binary or multi-classes will be available [30]. Support-vector machines are supervised learning models of machine learning with related classification models that conduct research for the study of classification and regression [31]. XGBoost is an algorithm for organized or tabular data that has lately dominated advanced machine learning and Kaggle events. XGBoost is an application of decision trees with gradient

boosts optimized for speed, and efficiency [32]. Random forests or random decision forests are a machine learning system for classification, regression, and other processes that work by creating at training phase a variety of decision trees and generating the category that would be the category mode or the specific trees information prior [33].

In deep learning, word embedding is the generic term for collecting language modeling, and function learning strategies where terms or words from the corpus are mapped to positive integer vectors [34]. An artificial recurrent neural network (RNN) model used in the area of deep learning is long short-term memory (LSTM). LSTM has reinforcement links, unlike normal recurrent neural networks. Not only does it handle single pieces of data, but also whole data series [35]. Two different approaches used for DL approaches. First one, word embedding layer along with dense vector. Second one, word embedding and lstm layer along with dense vector. In this paper, these models are implemented on all four datasets and the basket dataset.

IV. EXPERIMENTAL SETUP AND RESULTS

This section provides a detailed description of the experiments like environment setup, hyperparameter tuning, evaluation metrics, and corresponding results. It also provides a summary of the datasets used for the evaluation.

A. Experimental Setup

WWe have used python libraries for machine learning are NumPy, SciPy, Scikit-learn etc. The deep learning method is deployed on Tensorflow and Keras libraries. TensorFlow is a

TABLE II: Top 15 Spamming words and Sentences containing the spamming words from four different dataset

Dataset	Top 15 Spamming words	Sentence containing Top Spamming words
Trec-2007	dollar, transfer, org, price, items, office, now, save, email, said, click, please, viagra, time, online.	in exchange for all of the capital stock of sovs extreme has paid dollar 10k in cash and 15 million shares, it amp transfer the funds to the suppliers by means of western union money gram less your fee, www essentialmedicine org action many of our chapters have also posted photos on our flickr, lower the price of one of the most widely used aids drugs by 96 percent throughout sub-saharan africa, stability remaining uncorrupted products items shipping, the head office who will instruct you amp give advice regarding every new payment, go here now and get it, you have received this email because you have signed up to receive the cnn, he said after snapping pictures of the pill bottle sculpture at yale, alert preferences please click here, please visit the confirmation link below and fill out our short 30 second secure web, viagra if you have a problem getting or keeping an erection your sex life, the position offered is a part time job and will only require from you, repll cas is a well established online store.
Enron	dollar, company, will, percent, email, information, please, statements, may, one, now, money, time, business, 000.	we were compensated 3 000 dollars to distribute this report, i will do my best, believe that there could be a possible fit here for our company and our share holders, 100 percent all natural product that produces no side effects, you may purge your email address from our database, all information provided within this email pertaining to investing , stocks , securities must be understood as information provided and not investment advice, please copy and paste this link into your browser selftreatment, even after positive statements have been made regarding the above company, shares may be sold at any time, hot news flash today - this one is moving, get the magic blue pills now !, earn huge money quickly from home, this is how i made that money with the porn - chemisorb, the permanent fix to penis enlargement limited time offer, i intend to invest in real estate business your country with you as my partner, you have been pre - approved for a \$400 , 000 home loan at a 3 . 25 % fixed rate.
LingSpam	dollar, mail, order, report, email, will, address, free, program, money, send, name, list, one, business.	it is completely duplicatable ! by making sure you make thousands of dollars a week, use our server to send your own mail, order today ! only \$ 9 . 95 plus \$ 3 shipping and handling, for each report , send \$ 5 cash and a self-addressed stamped envelope, ! more features , more services (such as free email accounts!!), we will give you free of charge an original hand painted cel, 50 million e - mail addresses on a cd-rom , here 's a great directory for free and interesting internet sites, this company has been most effective for this program, this is a legitimate , legal , money-making opportunity, send us email with remove in the subject line, can you make one cent from each of theses names ?, the spider removes duplicates and saves the email list in a ready to send format, order any four videos and get the fifth one free, if you have been looking for a home-based business opportunity , this could be your lucky day.
Basket Dataset	dollar, percent, new, pills, email, one, price, transfer, top, now, please, mail, company, money, time, business.	get dollar 500 instantly when you register, the company guarantees to pay net 10 percent fee out of the amount of every payment, some additional openings for new employees we are glad to offer you, erection treatment pills , anti-depressant pills , weight loss , and more !, if you would prefer not to receive further emails fromthe savers club, we've had our eye on match point for almost one year now, offers the most money or offers the cheapest price per product, in fact it is my friend that advised me to urgently arrange to transfer out this money through another foreign beneficiary outside the country, we ' ve had our eye on match point for almost one year now, holds two wildcards that can be offered to other top piayers in the weeks leading up to the event, please hit your keyboard delete button now and please excuse the intrusion, are you in the mail order business ?, i will or the parent company will !, use hypnosis to make money !, a hard time finding places that will give you a credit card or a merchant card, own a corporation or own a business where you are under a duty to withhold from your employes paycheck.

repository of available and open deep learning applications. It could be used along with various functions but has a special emphasis on artificial neural network(ANN) training and

validation. Keras is an open-source framework that provides ANN with a Python interface. It acts as a framework for low-level libraries, written based on Python, such as TensorFlow

Fig. 2: WordClouds illustrating the significance of the top 100 spamming words extracted from each of the three different spam datasets and the Basket Dataset.

or Theano elevated neural network-based library. In simple words work using TensorFlow at backend. We have used GoogleColab for applying TensorFlow and Keras.

B. Evaluation Results

In this segment, We introduce several experimental findings obtained from various spam detection datasets and also our 'Basket' dataset.

Table II offers a list of top-15 overview sentences using the top-15 spamming Keywords that were derived from each of the data sets. From this table, it can be noted that most of the spamming words are important to the specific fields and may be helpful to visualize the corresponding corpus meaning.

Similarly, Figure 2 shows the wordclouds of the top 100 spamming words originating from different datasets. A word’s size and layouts are approximately equal to its significance for being spam email. To extract the spamming words, we only count the emails labeled as spam. It has been noticed from these wordclouds that all of the readily identifiable words are also correlated to the context of the individual datasets regardless of their larger font size.

We have split all four datasets into 80 percent training and 20 percent testing. The training part is used for model training. The test part evaluate the performance of the models. The numerical evaluation criteria used for machine learning in this paper include accuracy, precision, and recall metrics. Those are widely-used. For the deep learning part, we have calculated loss and accuracy.

After the deployment of ML/DL algorithms, the analytical method offers useful outcomes on each algorithm such as the accuracy of classification, precision, and recall, etc. We conduct four different classifiers algorithms in ml: Logistic regression, support vector machine, Xgboost and Random Forest.

In table III shows the machine learning approaches, which specifies dataset name, number of emails or datapoints for that dataset, number of features selected for training, ML model applied on that dataset, train loss, train accuracy, test loss, test accuracy, precision/recall for spam/ham (normal/legitimate) email.

In table IV shows the deep learning approaches, which

TABLE III: Four different machine learning models have applied on four renowned spam email datasets.

Dataset Name	Number of Email	Number of features	Model Name	Train LogLoss	Train Accuracy	Test LogLoss	Test Accuracy	Target	Precision for Test Data	Recall for Test Data
Track Dataset-2007	75392	7	Logistic Regression	0.576049	58.23%	0.578251	58.23%	Spam	0.737	0.944
Track Dataset-2007	75392	7	Support Vector Machine	2.28667	93.37%	14.46726	58.11%	Ham	0.873	0.531
Track Dataset-2007	75392	7	XGBoost	0.214022	93.37%	0.22707	93.17%	Spam	0.582	0.995
Track Dataset-2007	75392	7	Random Forest	9.992001e-16	100%	2.3432578	93.21%	Ham	0.371	0.004
Enron email dataset	33062	2	Logistic Regression	0.42102	82.81%	0.426375	52.60%	Spam	0.896	0.998
Enron email dataset	33062	2	Support Vector Machine	9.9920e-16	100%	5.275130	84.72%	Ham	0.997	0.839
Enron email dataset	33062	2	XGBoost	0.264829	91.06%	0.295188	88.46%	Spam	0.896	1.00
Enron email dataset	33062	2	Random Forest	9.99200e-16	100%	2.663691	92.28%	Ham	0.896	0.838
PU Dataset	6961	2	Logistic Regression	0.577633	58.83%	0.567477	58.50%	Spam	0.880	0.861
PU Dataset	6961	2	Support Vector Machine	9.9920e-16	100%	13.73617	60.22%	Legitimate	0.864	0.882
PU Dataset	6961	2	XGBoost	0.36291	88.07%	0.45524	80.40%	Spam	0.889	0.810
PU Dataset	6961	2	Random Forest	9.99200e-16	100%	6.471420	81.26%	Legitimate	0.825	0.898
Lingspam Dataset	11314	2	Logistic Regression	0.13184	83.471%	0.133466	83.473%	Spam	0.894	0.874
Lingspam Dataset	11314	2	Support Vector Machine	9.99200e-16	100%	3.32720	90.36%	Legitimate	0.876	0.896
Lingspam Dataset	11314	2	XGBoost	0.679903	98.85%	0.68039	95.89%	Spam	0.919	0.928
Lingspam Dataset	11314	2	Random Forest	9.99200e-16	100%	1.358353	96.06%	Legitimate	0.927	0.918
Basket Dataset	126246	2	Logistic Regression	0.63658	54.43%	0.636478	54.43%	Spam	0.855	0.311
Basket Dataset	126246	2	Support Vector Machine	10.5743	69.38%	13.378068	61.26%	Legitimate	0.656	0.961
Basket Dataset	126246	2	XGBoost	0.556454	69.38%	0.565057	68.48%	Spam	0.750	0.913
Basket Dataset	126246	2	Random Forest	1.036564	96.99%	10.420631	69.82%	Legitimate	0.595	0.979
								Spam	0.805	0.707
								Legitimate	0.804	0.875
								Spam	0.829	0.701
								Legitimate	0.840	0.894
								Spam	0.960	0.703
								Legitimate	0.944	0.994
								Spam	0.959	0.436
								Legitimate	0.899	0.996
								Spam	0.961	0.783
								Legitimate	0.959	0.994
								Spam	0.990	0.770
								Legitimate	0.956	0.998
								Spam	0.610	0.866
								Legitimate	0.744	0.414
								Spam	0.578	0.921
								Legitimate	0.773	0.285
								Spam	0.637	0.903
								Legitimate	0.815	0.453
								Spam	0.661	0.851
								Legitimate	0.772	0.537

TABLE IV: Deep Neural Networks have applied on our selected four renowned spam email datasets.

Dataset Name	Number of Email	Deep Learning Approach	Train Loss	Train Accuracy	Test Loss	Test Accuracy
Trecspam dataset-2007	75419	WordEmbedding	2.2010e-09	100%	0.0160	99.80%
Trecspam dataset-2007	75419	WordEmbedding+LSTM	8.5344e-04	99.96%	0.0097	99.73%
Enron email dataset	18180	WordEmbedding	3.5710e-06	100%	0.0759	98.33%
Enron email dataset	18180	WordEmbedding+LSTM	9.5358e-04	99.99%	0.0904	98.22%
PU dataset	7101	WordEmbedding	9.3088e-05	100%	0.1430	96.53%
PU dataset	7101	WordEmbedding+LSTM	0.0376	99.26%	0.1921	95.1%
Lingspam dataset	11572	WordEmbedding	3.3348e-06	100%	8.7080e-05	100%
Lingspam dataset	11572	WordEmbedding+LSTM	4.8243e-06	100%	0.0033	99.94%
Basket Dataset	127808	WordEmbedding	3.0199e-05	100%	0.1757	98.53%
Basket Dataset	127808	WordEmbedding+LSTM	0.0012	99.95%	0.0547	99.05%

specifies dataset name, Number of email, Deep learning Model, train loss, train accuracy, test loss, and test accuracy.

In figure 3 shows the log-loss comparison of four different machine learnign models on four different datasets.

C. Discussion

First of all we extracted different features from each datasets. In these case we found subject and body for Enrom, Ling-spam, and PU dataset. But Trec dataset of 2007 offers approximately 80 different features. We selected only those features that carry enough information for classifying among spam and legitimate emails. In table III shows that we took 2 features for Enrom, PU, and Lingspam. Where's we found 7 features from trec-2007 countable for our research. We merge all four datasets in a combined one called 'Basket Dataset'. Basket dataset is also treated as a particular dataset in our study. In Basket dataset subject and body of the four datasets considered only. After successful feature extraction we pre-process the data so that no noise hampered our model to learn detail knowledge. Feature engineering is perform on our

Fig. 3: Log loss comparison among different datasets in machine Learning approaches.

pre-processed dataset. In ML we perform average word2vec with tf-idf, in dl we perform one-hot encoding, tokenization interchangeably.

After performing all pre-processing steps, machine/deep learning models deployed on our data. In table III we summarize the details of ML outcomes. The performance of logistic regression quite well for ling-spam dataset. Where support vector machine overfitted for all four datasets. XGBoost overfitted for PU dataset but perform well for Trec,enrom, and Lingspam. Random Forest's performance is quite good but overfitted for all except lingspam. Based on performance metrics XGBoost perform well than others. From figure 3 and figure III, we found Logistic regression loss ranges from 0.13 to 0.57, XGBoost loss ranges from 0.22 to 0.68 for different datasets. Support vector machine and random forest's train loss is low but test loss is dramatically high. For trec-spam dataset ML model performs well to classify spam email. But, for other dataset ML model performs well to classify legitimate email.

So, considering precision, recall, accuracy metrics and loss function, XGBoost is the best machine learning model for all four datasets.

In table IV shows the summary of performed deep learning approaches. Loss is comparatively low for all approaches. Training accuracy is over 99% Test accuracy is quite good for all datasets, which is over 95% No overfitting encountered in any cases. It performed well even for 'Basket dataset'. Word-Embedding (along with dense layer) accuracy is 98.53% and loss 0.175 for Basket dataset. Moreover, Basket dataset's LSTM and word-embedding approach accuracy is 99.05% and loss 0.0547. But machine learning models loss are quite high and train-test accuracy surprisingly low on 'Basket Dataset'. Word-embedding layer perform slightly better than word embedding and LSTM layer in case of Trec-2007, Enrom, PU and Lingspam datasets. But for our basket dataset word embedding and LSTM layer performed well.

After observing all result deep learning approaches performed state of the art, not only for each dataset but also for our basket dataset. Our basket dataset helps us to create a more robust and multi-dimensional model. As all four datasets are from different domains and covered different types of spam email, our model could perform in wide range of area.

V. CONCLUSION

Spam or unsolicited junk mail is both harmful and annoying. In this paper, we have developed a pipeline to identify the junk mails from the legal one. We have applied machine learning and deep learning algorithm for four individual datasets. We have found that the XGBoost is the best machine learning model for all four datasets. In deep learning, the Word-embedding layer performed better. We have also used 'Basket' dataset and the best result we got by implementing LSTM. Finally, hybrid systems seem to be the most effective method for creating a reliable anti-spam filter today.

REFERENCES

- [1] C. Bird, A. Gourley, P. Devanbu, M. Gertz, and A. Swaminathan, "Mining email social networks," in *Proceedings of the 2006 international workshop on Mining software repositories*, 2006, pp. 137–143.
- [2] S. Silberfenig and E. Karich, "Combination cellular telephone, sound storage device, and email communication device," Apr. 25 2006, uS Patent 7,035,666.
- [3] R. Barnes, B. Schneier, C. Jennings, T. Hardie, B. Trammel, C. Huitema, and D. Borkman, "Confidentiality in the face of pervasive surveillance: A threat model and problem statement," 2015.
- [4] S. Whittaker and C. Sidner, "Email overload: exploring personal information management of email," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, 1996, pp. 276–283.
- [5] R. Michalski, S. Palus, and P. Kazienko, "Matching organizational structure and social network extracted from email communication," in *International Conference on Business Information Systems*. Springer, 2011, pp. 197–206.
- [6] X. Hu, B. Li, Y. Zhang, C. Zhou, and H. Ma, "Detecting compromised email accounts from the perspective of graph topology," in *Proceedings of the 11th International Conference on Future Internet Technologies*, 2016, pp. 76–82.
- [7] R. Lieb, "Make spammers pay before you do," *The ClickZ Network*, 2002.
- [8] L. Zhuang, J. Dunagan, D. R. Simon, H. J. Wang, I. Osipkov, and J. D. Tygar, "Characterizing botnets from email spam records," *LEET*, vol. 8, pp. 1–9, 2008.
- [9] V. Rajan, M. Wegman, R. Segal, J. Crawford, J. Kephart, and S. Hershkop, "Detecting spam email using multiple spam classifiers," Jul. 6 2006, uS Patent App. 11/029,069.
- [10] A. K. Sharma, S. K. Prajapat, and M. Aslam, "A comparative study between naïve bayes and neural network (mlp) classifier for spam email detection," *Int. J. Comput. Appl.*, 2014.
- [11] C. Palanisamy, T. Kumaresan, and S. Varalakshmi, "Combined techniques for detecting email spam using negative selection and particle swarm optimization," *Int. J. Adv. Res. Trends Eng. Technol.*, vol. 3, 2016.
- [12] A. Tyagi, "Content based spam classification-a deep learning approach," Master's thesis, Graduate Studies, 2016.
- [13] B. Zhou, Y. Yao, and J. Luo, "Cost-sensitive three-way email spam filtering," *Journal of Intelligent Information Systems*, vol. 42, no. 1, pp. 19–45, 2014.
- [14] Z. Zhong, L. Ramaswamy, and K. Li, "Alpacas: A large-scale privacy-aware collaborative anti-spam system," in *IEEE INFOCOM 2008-The 27th Conference on Computer Communications*. IEEE, 2008, pp. 556–564.
- [15] C. H. Li and J. X. Huang, "Spam filtering using semantic similarity approach and adaptive bpnn," *Neurocomputing*, vol. 92, pp. 88–97, 2012.

- [16] S. Dhavale, "C-asft: Convolutional neural networks-based anti-spam filtering," in *Proceeding of International Conference on Computational Science and Applications: ICCSA 2019*. Springer Nature, 2020, p. 49.
- [17] S. S. Roy, A. Sinha, R. Roy, C. Barna, and P. Samui, "Spam email detection using deep support vector machine, support vector machine and artificial neural network," in *International Workshop Soft Computing Applications*. Springer, 2016, pp. 162–174.
- [18] A. Zamir, H. U. Khan, W. Mehmood, T. Iqbal, and A. U. Akram, "A feature-centric spam email detection model using diverse supervised machine learning algorithms," *The Electronic Library*, 2020.
- [19] S. Seth and S. Biswas, "Multimodal spam classification using deep learning techniques," in *2017 13th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*. IEEE, 2017, pp. 346–349.
- [20] A. N. Soni, "Spam-e-mail-detection-using-advanced-deep-convolution-neuralnetwork-algorithms," *JOURNAL FOR INNOVATIVE DEVELOPMENT IN PHARMACEUTICAL AND TECHNICAL SCIENCE*, vol. 2, no. 5, pp. 74–80, 2019.
- [21] C. Wang, Q. Li, T.-y. Ren, X.-h. Wang, and G.-x. Guo, "High efficiency spam filtering: A manifold learning-based approach," *Mathematical Problems in Engineering*, vol. 2021, 2021.
- [22] S. Srinivasan, V. Ravi, M. Alazab, S. Ketha, A.-Z. Ala'M, and S. K. Padannayil, "Spam emails detection based on distributed word embedding with deep learning," in *Machine Intelligence and Big Data Analytics for Cybersecurity Applications*. Springer, 2021, pp. 161–189.
- [23] M. E. Manaa, A. J. Obaid, and M. H. Dosh, "Unsupervised approach for email spam filtering using data mining," 2021.
- [30] D. W. Hosmer Jr, S. Lemeshow, and R. X. Sturdivant, *Applied logistic regression*. John Wiley & Sons, 2013, vol. 398.
- [24] G. Cormack, "Trec 2007 spam track overview," 01 2007.
- [25] B. Klimt and Y. Yang, "The enron corpus: A new dataset for email classification research," in *European Conference on Machine Learning*. Springer, 2004, pp. 217–226.
- [26] "Pu123acorpora dataset," <http://www.aueb.gr/users/ion/data/PU123ACorpora.tar.gz>.
- [27] "Ling spam public dataset," http://www.aueb.gr/users/ion/data/lingspam_public.tar.gz.
- [28] F. Heimerl, S. Lohmann, S. Lange, and T. Ertl, "Word cloud explorer: Text analytics based on word clouds," in *2014 47th Hawaii International Conference on System Sciences*. IEEE, 2014, pp. 1833–1842.
- [29] M. Abulaish, M. I. H. Showrov, and M. Fazil, "A layered approach for summarization and context learning from microblogging data," in *Proceedings of the 20th International Conference on Information Integration and Web-based Applications & Services*, 2018, pp. 70–78.
- [31] L. Auria and R. A. Moro, "Support vector machines (svm) as a technique for solvency analysis," 2008.
- [32] H. Zheng, J. Yuan, and L. Chen, "Short-term load forecasting using emd-lstm neural networks with a xgboost algorithm for feature importance evaluation," *Energies*, vol. 10, no. 8, p. 1168, 2017.
- [33] G. Biau, "Analysis of a random forests model," *The Journal of Machine Learning Research*, vol. 13, no. 1, pp. 1063–1095, 2012.
- [34] J. Bian, B. Gao, and T.-Y. Liu, "Knowledge-powered deep learning for word embedding," in *Joint European conference on machine learning and knowledge discovery in databases*. Springer, 2014, pp. 132–148.
- [35] S. Bouktif, A. Fiaz, A. Ouni, and M. A. Serhani, "Optimal deep learning lstm model for electric load forecasting using feature selection and genetic algorithm: Comparison with machine learning approaches," *Energies*, vol. 11, no. 7, p. 1636, 2018.