Evaluation of Classification Algorithms for Effective Spam Email Detection using Spam Email Dataset

Kalyani Shivaji Ubale[1], Kamini Ashutosh Shirsath[2]
[1]Research Scholar,
Computer Engineering Department,
K K Wagh Institute of Research and Technology,
SPPU, Pune, Nashik, India
E-mail: kalyaniubale110@gmail.com

[2]Computer Engineering Department,
Sandip Institute of Engineering and Management,
SPPU, Pune, Nashik, India
E-mail: kamini.nalavade@live.com

Abstract— **The ever-changing digital landscape of today has led to an unsettling increase in spam emails, which pose a serious risk to cyber security due to the nature of email communication. Email marketing campaigns, user interaction, and important notifications are all supported by a large number of organizations, which has led to calls for spam emails to be responded to right away. Spam emails are made up of a wide range of uninvited and unwelcome communications that are frequently disguised using social engineering, malware distribution, and phishing techniques. The sophistication of spamming techniques has made conventional filtering strategies ineffective, requiring a more thorough comprehension of the dynamic nature of spam. Researchers may examine and predict the complex patterns and properties of spam emails with the help of free spam email databases, which are an essential resource. These datasets enable the creation and validation of complex data mining algorithms for spam identification and prevention by offering a realistic and dynamic depiction of the always changing spam landscape. The datasets present a wide range of spam samples, which ensures a thorough comprehension of the strategies used by spammers in different situations. In this paper, we attempt to present analysis of various classifiers on the complex datasets of spam emails Additionally, we provided a comprehensive analysis of all the datasets that are publically accessible by contrasting them based on a range of criteria, including the percentage of spam emails, the total number of occurrences, the number of attributes taken into consideration, the type of feature, etc.**

Index Terms— **Spam, Spam Filtering Method, Naive Bayes, K-Nearest Neighbors, SVM, Logistic regression, Decision tree, Random forest;**

1. INTRODUCTION

In the current era of digitalization, where communication is primarily accomplished through electronic channels, uninvited and undesired emails—also referred to as spam—are a typical occurrence. Worldwide, spam emails flood inboxes with unsolicited content, frauds, and promotions, clogging up both personal and professional accounts. It is important for organizations to protect their networks and reputation as well as for individuals to maintain effective communication channels to comprehend the effects of spam emails. Currently, almost 85% of emails and texts that users receive are spam. The total number of emails sent every day is increasing each year. The total growth was observed from 269 to 333.2 billion, between the year 2017 and 2022, which means that over 64 billion more emails were sent every day compared to the last 5 years. The people say that, the most common medium they receive spam through is email, with almost half (49%) saying they receive spam via emails most often [1]. The volume of data makes it impractical to manually analyze spam communications. Using machine learning techniques, the most accurate spam classification may be accomplished. Junk email, sometimes known as spam, is unsolicited communication delivered by spammers by email [2]. The following are the primary categories of unsolicited emails: advertisements; Nigerian spam, which is spam sent by con artists attempting to extract money from letter recipients; Phishing is the practice of building a fake website that looks like it belongs to a reputable company in an effort to electronically gain sensitive or private information from visitors (typically with the intention of stealing it) [3]. This message lacks a valid return address.

The carbon footprint of one spam email is around 0.03g CO2e. Thus, the amount of spam emails sent in 2021 may have contributed to the release of 4.5 tons of CO2e. Different kinds of spam emails are sent. Marketing, advertising, sexual content, financial matters, scams, and fraud are among the most prevalent categories of spam. According to a poll conducted in October 2021, the top 10 nations that send the highest percentage of spam emails are the United States, China, Russia, Brazil, India, Germany, Czech Republic, Poland, Bulgaria, and the United Kingdom [2]. To gain a better understanding of the scope of the spam problem caused by AT&T and Lucent subdomains over a six-month period starting at the end of April 1997, a case study was conducted.

Due to spammers' evolving strategies, a preliminary review of data gathered at the end of six months revealed a decline in filter performance [3]. Additionally, it was noted that, in the past, spammers would typically include messages in the body of their emails; this practice was termed as "traditional spamming." However, throughout the past ten years, email services have faced a new challenge in the form of image spam, which has evolved into a sophisticated form of spam since it makes the message more engaging for the user and makes it difficult for filters to identify as spam. Image spam filters use various techniques on their filters to be more effective because image spam is rich in information and contains a range of data.

Thus, there is a growing demand for trustworthy anti-spam filters due to the volume of unsolicited bulk emails. Thus far, these kinds of filters have primarily relied on manually created keyword patterns that yield subpar results. Using the publicly accessible email corpus, a comprehensive analysis of the Naïve

Bayesian filter's performance is conducted, helping to establish industry standards [4]. The Reuters corpus, which is a publicly accessible collection of human categorized documents, has been beneficial to the field of text categorization research (Lewis, 1992). The same have served as reference points. One of the issues with having a standard email corpus publicly accessible is that it is difficult to provide resources for anti-spam filtering similar to this since users' incoming emails cannot be made public without breaking

Combining spam messages with ones taken from public mailing list archives that are free of spam is one of the most popular approaches to solving this kind of issue. In that perspective, we test Sahami et al.'s methodology using a combination of spam and communications submitted through the moderated (i.e., spam-free) Linguist list, which discusses the field and science of linguistics. The final corpus, known as Ling-Spam, is released to the public so that it can be used as a benchmark by others [5,6]. Making spam messages public is not a problem because spam is essentially already public knowledge because it is sent out blindly to a vast number of recipients. However, it is generally more difficult to publish genuine messages online without invading the privacy of those who send and receive them.

In an attempt to provide a comprehensive analysis, we attempt to suggest a few of the most well-known publicly accessible spam email corpora. In order to help academics working with the spam data use the various email spam datasets more effectively, we attempted to emphasize in this paper their fundamental characteristics along with their benefits and drawbacks. Additionally, we suggested other dataset qualities by taking into account the characteristics associated with spam emails.

2. Literature Survey

The attempts to enact laws prohibiting spam emails have not had much of an impact. Creating tools to assist recipients in recognizing or deleting spam mail automatically is a more efficient approach. These devices are known as anti-spam filters, and their features range from content-based filters to blacklists of known spammers. A fresh benchmark corpus was created, containing a combination of spam and communications from moderated (thus, spam-free) email lists. The corpus is released to the public so that other academics can use it as a standard. An extensive analysis of the Naive Bayesian technique, which was employed in (Sahami et al. 1998), was carried out using this corpus [4]. An extensive analysis on a publicly available corpus was carried out, which helped establish common standards.

A comprehensive analysis was carried out on a publicly available corpus, which helped establish common standards. Moreover, previously unexplored topics such as the impact of stop lists, lemmatization, training corpus size, and attribute set size on the filter's performance were examined. In order for the Naive Bayesian anti-spam filter to be practical in practice, more safety nets must be included, it was determined after implementing suitable cost-sensitive assessment measures[5]. In the context of anti-spam filtering, a unique cost-sensitive application of text categorization, a method for stacking classifiers—known as stacked generalization—was assessed. They demonstrated that stacking can increase the effectiveness of automatically generated anti-spam filters and that these filters can be applied in real-world scenarios using a publicly available corpus known as Lingspam [6].

An assessment of memory-based learning in relation to anti-spam filtering is presented in this research. It also suggested a unique, cost-sensitive use of text categorization in an effort to recognize automatically the bulk of unwanted commercial emails that arrive in inboxes. Using a publicly available corpus, a comprehensive study of memory-based anti-spam filter effectiveness is conducted with an emphasis on anti-spam filtering for mailing lists. A variety of attribute and distance-weighting approaches are examined, together with studies on the impact of neighborhood, attribute set, and training corpus sizes [7]. Four learning algorithms—Naive Bayes, Flexible Bayes, LogitBoost, and Support Vector Machines—as well as four datasets made from various users' mailboxes were the subjects of an inquiry. We talked about the worst-case computational complexity numbers.

It is carried out a study on the impact of employing attributes that reflect token sequences rather than individual tokens on classification accuracy. It was also investigated how the size of the training set and attribute affected things within a budget-conscious framework [8]. Five supervised learning techniques are evaluated and presented in relation to statistical spam filtering. Using cost-sensitive metrics, the effects of various feature trimming techniques and feature set sizes on the performance of every learner are also investigated. It was found that from classifier to classifier, the importance of feature selection differs significantly. Support vector machines, AdaBoost, and maximum entropy models, in particular, perform well in the evaluation that is conducted and have comparable characteristics: they are not sensitive to feature selection strategies and can be easily scaled to very high feature dimension, and good performances across different datasets

The four spam corpora used for the evaluation were PU1, Ling, SA, and ZH1 [9]. A comparison was done of SVM, NB, boosted trees and stacking algorithms on the benchmark spam filtering corpora LingSpam and PU1. Two disjoint natural feature sets must define the datasets in order to do conventional semi-supervised co-training. The majority of datasets only contain one set of features, which restricts co-training's application. It was observed that if there is high data redundancy as in the domain of spam e-mail filtering, co-training with random feature split is as competitive as co-training with natural feature split[10].

A comparison between the linear Support Vector Machine used to automatically filter spam emails and seven distinct Naive Bayes classifier versions is shown. Six renowned, sizable, publicly accessible databases—collectively referred to as the Enron corpus—were used in empirical studies. The outcomes showed that Boolean NB, MN, and linear SVM If you want to automatically filter spam, Boolean NB and Basic NB are your best options. Nonetheless, SVM demonstrated the highest average performance across all examined databases, exhibiting an accuracy rate exceeding 90% across every corpus examined. After dimensionality reduction using eight well-known term-selection strategies with differing degrees of popularity, the performance of seven distinct Naive Bayes spam filters deployed to categorize messages from six well-known, actual, public, and huge e-mail data sets was compared. The comparison of the performance achieved by seven different Naive Bayes spam filters applied to classify messages from six well-known, real, public and large e-mail data sets, after a step of dimensionality reduction employed by eight popular term-selection techniques varying the number of selected terms was performed [11,12,18].

3. Analysis of Dataset

The different publicly available email spam corpuses are described in this section. The spam email databases provide a wide range of data, including picture, text, and phishing email data. Recently, a significant rise in spam emails pertaining to all of the previously listed categories of spam data has been noted. Making spam messages public is not a problem because spam is essentially already public knowledge because it is sent out blindly to a vast number of recipients. However, legitimate messages often cannot be distributed without infringing upon the privacy of both senders and recipients. Using authentic communications gathered from publicly accessible newsgroups or mailing lists with public archives is one approach to get around privacy issues.

The field of cybersecurity and spam detection has evolved tremendously as a result of the availability of publicly available spam emails for study. We have examined the many advantages and ramifications of using these corpora throughout this review paper, highlighting their critical influence in molding the creation of efficient spam filtering methods and deepening our comprehension of spamming behavior. The fact that publicly accessible spam email corpora encourage creativity and teamwork among researchers is one of their main benefits. These corpora facilitate the systematic evaluation and comparison of spam detection algorithms, allowing academics to identify new approaches and best practices by giving them access to standardized datasets with ground truth labels. Furthermore, this corpora's open nature promotes repeatability and transparency. The table I shows comparison of various spam email datasets and the commonly considered features for spam filtering.

Furthermore, publicly accessible spam email corpora are priceless tools for researching the constantly changing techniques and approaches used by spammers. Researchers can learn more about the underlying patterns and trends that motivate spamming activity by examining large-scale datasets from a variety of sources and historical periods. The table II shows publicly available email spam corpus and their data characteristics and the analysis of the number features that are commonly considered for spam detection.

TABLE I: COMPARING VARIOUS DATASETS WITH THEIR FEATURES

Dataset	Source	Size	Label (Spam/ Ham)	Features
Spam Archive	Research dataset	15.4 MB	Binary	Text content, sender, recipient, subject, date/time, attachments
Spam email	Research dataset	Varies	Binary	- Subject line - Body text - Sender email address - Recipient email address
Spam base	UCI machine learning repository	0.4 MB	Binary	- Word frequency features (e.g., 'make', 'money', 'free'), character frequency, special characters, capitalization
Spam assassin	Spam Assassin public corpus	3.3 GB	Binary	- Email text (both body and headers) - Metadata (e.g., date, sender, subject)
Ling spam	Lingspam dataset	20 MB	Binary	- Bag of words features - Metadata (e.g., subject line, sender, date)
PU corpus	Research dataset	0.4 GB	Binary	- Bag of words features - Metadata (e.g., subject line, sender, date)
Phishing corpus	Research dataset	Varies	Binary	- HTML content features - Metadata (e.g., URL, domain)
Zh1	Research dataset	Varies	Binary	- Bag of words features - Metadata (e.g., subject line, sender, date)
Gen spam	Research dataset	Varies	Spam	Text content, sender, recipient, subject, date/time, attachments
Princeton spam image	Princeton university	Varies	Spam	Image pixels, image metadata, sender, recipient, subject, date/time
Dredze image spam	Johns Hopkins university	Varies	Spam	Image pixels, image metadata, sender, recipient, subject, date/time
Hunter	Public sources	Varies	Spam	Text content, sender, recipient, subject, date/time, attachments
Enron spam	Enron Corporation	12.8 MB	Binary	- Email text (both body and headers) - Metadata (e.g., subject line, sender, date)
Trec	TREC public spam corpora	Varies	Binary	- Email text (both body and headers) - Metadata (e.g., date, sender, subject)

TABLE II: PUBLICLY AVAILABLE EMAIL SPAM CORPUS

Dataset Name	Number of messages		Rate of spam	Year of creation	References	Dataset characteristic	Associated tasks	Feature type	Features
	Spam	Non-spam							
Spam archive	15090	0	100%	1998	Almeida and yamakami	Multivariate	Classification	Integer	Varies
Spambase	1813	2788	39%	1999	Sakkis et al	Multivariate	Classification	Integer, Real	57
Lingspam	481	2412	17%	2000	Sakkis et al	Multivariate	Classification	Char, Integer	50
PU1	481	618	44%	2000	Attar et al	Multivariate	Classification	Integer	Varies
Spamassassin	1897	4150	31%	2002	Apache spam-assassin	Multivariate	Classification	Integer	100
PU2	142	579	20%	2003	Zhang et al	Multivariate	Classification	Integer	Varies
PU3	1826	2313	44%	2003	Zhang et al	Multivariate	Classification	Integer	Varies
PUA	571	571	50%	2003	Zhang et al	Multivariate	Classification	Integer	Varies
Zh1	1205	428	74%	2004	Zhang et al	Multivariate	Classification	Integer, Char	3000
Trec 2005	52,790	39,399	57%	2005	Androutsopoulos et al	Multivariate	Prediction	Integer, Char, Real	Varies
Phishing corpus	415	0	100%	2005	Abu-nimeh et al	Multivariate	Classification	Integer, Real	43
Enron-spam	20170	16545	55%	2006	Koprinska et al	Multivariate	Classification	Integer, Real	375
Trec 2006	24,912	12910	66%	2006	Androutsopoulos et al	Multivariate	Prediction	Integer, Char, Real	Varies
Trec 2007	50,199	25,220	67%	2007	Debarr and Wechsler	Multivariate	Prediction	Integer, Char, Real	135
Princeton spam image Benchmark	1071	0	100%	2007	Wang et al	Multivariate	Classification	Integer, Char, Real	50
Dredze image spam Dataset	3297	2021	62%	2007	Dredze,gevaryahu and elias-bachrach	Multivariate	Classification	Integer, Char, Real	23
Hunter	928	810	53%	2008	Gao et al	Multivariate	Classification	Integer, Char, Real	24
Spam email	1378	2949	32%	2010	Csmining group	Multivariate	Classification	Integer, Real	64

This knowledge can then be used to inform the creation of more reliable and adaptable spam detection systems. Moreover, the incorporation of metadata, such as email headers and sender information, allows scholars to explore the technological and social dimensions of spam activity, providing insights into the mechanisms and incentives underlying spam campaigns. Additionally, the accessibility of spam email corpora has promoted cross-disciplinary cooperation between academics in computer science, linguistics, psychology, and other disciplines.

4. PROPOSED SYSTEM

These days, the most widely used spam filtering techniques are as follows;

1. Systems that require confirmation. In order to guarantee that the original message is sent, the sender is invited to take action; if not, the message is deemed undelivered.
2. Utilizing provisional mailing addresses. When there are a lot of arriving letters, the user updates the address.
3. A blacklist. Upon receiving an incoming message, the spam filter verifies if the sender's IP address or email address is listed on the blacklist. If it is, the message is deemed spam and is discarded [8].
4. Whitelist. The principle of operation is the same as in the method with black lists, but the check is made for the absence of the sending IP address in the black list of the mail server.
5. Spam recognition based on signatures. A signature is an image or characteristic of an email message. For each new message, its signature is calculated and compared with the database, which stores the characteristics of messages previously classified as spam. If the message signature matches one of the database records, the message is considered as spam.
6. Linguistic heuristics. Search in the body of the message for keywords and phrases that allow attributing this message to spam.

Machine learning is a broad field within artificial intelligence that focuses on developing algorithms that are capable of learning from examples and making predictions. The most widely-liked machine learning methods for categorizing spam as well as current spam detection strategies will be examined in the following section.

We will start by outlining the key phases of the machine learning process. First, the analysis stage: in this phase, processed and analyzed data are used to identify patterns. Second, train stage: using the acquired data, machine learning models are applied. The choice of hyper parameters has the potential to enhance the models' quality. Testing comes next: On unutilized data, machine learning models are tested. The model can be assessed with a variety of indicators. Application, or implementing the best model, is the final step.

There are six widely spread classification algorithms in machine learning which were selected: Logistic regression, KNearest Neighbors, Ada Boost, Naive Bayes, Gradient Foresting, Random forest.

A probabilistic algorithm that is good at classifying spam is called Naïve Bayes. It reduces a multidimensional problem to a collection of univariate problems, ignoring potential relationships or connections among inputs, which is why it is referred to as "naive" [9]. The following are the drawbacks of using this algorithm to process spam emails: The quality of classification will suffer if a term in the letter appears that has never been in the training sample.

KNN is a metric classification approach in which distances between items are computed and objects are represented as points in space. After that, it moves into a learning phase in which training data points are repeatedly allocated to a cluster whose center is closest in distance [10]. One can optimize the tuning of the algorithm's input parameter, k. The selected k value determines the accuracy of the classification. The training sample contains k of the categorized object's closest neighbors. An object is a member of the class that its k closest neighbors share the most of. The algorithm does not function properly with a high number of characteristics and is unstable to outliers.

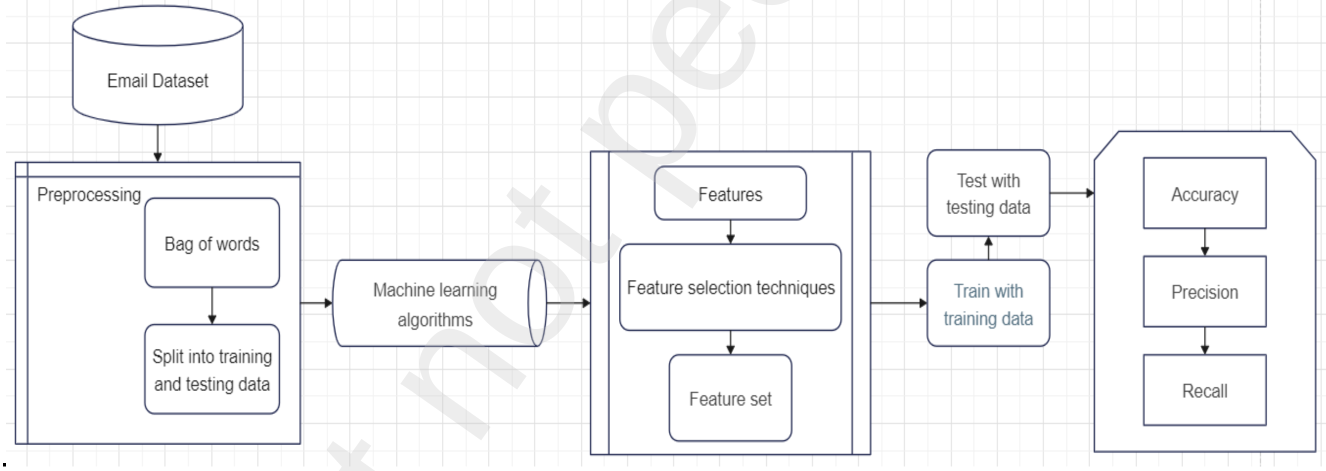

FIGURE I: SPAM DETECTION BLOCK DIAGRAM

An appropriate analytical technique for modeling the data and elucidating the correlation between the explanatory variables and the binary answer variable is logistic regression. The chance of allocating a value to a certain class is the outcome, and it can only take values between 0 and 1 [11].

A prediction method called Random Forest makes advantage of the tree-building concept. The combined effect of multiple trees, or a forest, improves each tree's capacity for prediction independently. The programmer builds a number of decision trees during the training phase, which are subsequently utilized for class prediction. This is accomplished by taking into account the voted classes of each individual tree, with the output being the class with the highest vote [3].

5. Experimentation and Results

The main stages of creating a classifier are: 1) Data collection; 2) Pre-processing and text cleaning; 3) Learning and obtaining prediction results. The main task of the algorithm is to find patterns.

The Lingspam dataset was used to compare the performance of the Logistic regression, Neighbors classifier, Ada boost classifier, Naïve bayes, Gradient boosting classifier and Random forest classifier. The benchmark corpus lingspam is a mixture of spam messages and messages received via the Linguist list, a moderated mailing list about the profession and science of linguistics. The corpus contains 10 directories. It consists of 2893 messages among which 2412 are Linguist messages, obtained by randomly downloading digests from the list's archives, breaking the digests into their messages, and removing text added by the list's server. 481 spam messages, received by the first author. Attachments, HTML tags, and duplicate spam messages received on the same day is not included. Spam messages are 16.6% of the corpus. The Figure I represents the process that is followed to implement the model.

Data processing and algorithm creation process:

Preparing the data is the initial step in the data processing and algorithm creation process. The data must first be cleaned up of gaps, duplication, and undefined values. Even if many models permit the inclusion of gaps in the sets and undefined data, it is preferable to erase them before executing simple changes in order to reduce the likelihood of errors and enhance the quality of the classification.

The Table III shows the composition of lingspam dataset.

TABLE III: LINGSPAM DATASET

Dataset	**Spam**	**Non-spam**	**Rate of spam**	**Size**	**Year of creation**	**Features**
Lingspam	481	2412	16.6%	20 MB	2000	50

Table IV present the performance achieved by each spam classifier algorithm for Lingspam dataset. For the lingspam data, all the classifiers evaluated only 14 randomly selected features at a time to find the best feature for each of its 50 features.

TABLE IV: PERFORMANCE COMPARISON FOR THE METHODS EVALUATED

Algorithm	**Accuracy**	**Precision**	**Recall**
Logistic Regression	0.972	0.977	0.972
K Neighbors	0.955	0.955	0.955
Ada Boost	0.986	0.986	0.986
Naïve Bayes	0.872	0.970	0.872
Gradient boosting	0.968	0.972	0.968
Random forest	0.982	0.984	0.982

In this experiment, Kaggle platform was used for programming. The dataset was divided into two parts; training and testing data. The main aim of this learning is to train a classifier with the lingspam dataset and then predict the spam email with the testing lingspam dataset. Experimental results for the lingspam data, shows that the Adaboost and Random forest method significantly improves the performance of the spam filter compared to the other classifiers like logistic regression, Neighbors, Naïve bayes and Gradient boosting algorithms. The Fig II. Shows comparative analysis of the various spam classifier algorithms.

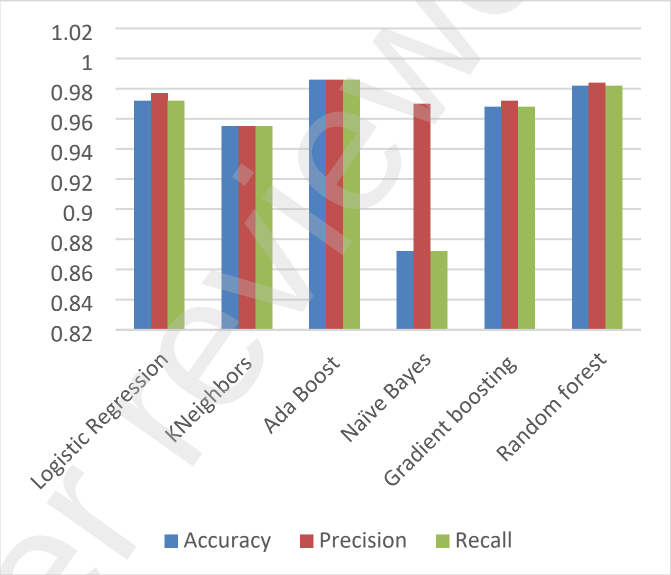

FIG II: COMPARATIVE ANALYSIS OF THE VARIOUS SPAM CLASSIFIER ALGORITHMS

6. Conclusion

In conclusion, we can state that spam is an issue that bothers every user, which is why we have done study on the subject. To create spam protection, there are multiple approaches. Spam message classification is highly accurate thanks to machine learning algorithms. In this study, training was conducted using an available lingspam dataset. 481 spam mails and 2412 non-spam samples make up the data. These outcomes were attained using a variety of machine learning models, including Random Forest, Naive Bayes, K-Nearest Neighbors, Ada Boost, and Logistic Regression. In this work, we observe that, due to their highest accuracy, the most effective algorithms for spam filtering are Logistic Regression and Random Forest. It is now at 99%. It has reached 99%. The results can be used to create a more intelligent spam detection classifier by combining algorithms or filtering methods.

7. References

[1] Cai, Robert, "Spam statistics(2024): New data on junk email, AI scams and phishing", online blog, 19 Oct 2023.

[2] Jyothiikaa Moorthy, "23 email spam statistics to know in 2024", online blog, 8 Aug 2023.

[3] L. F. Cranor, B. A. Lamacchia, "Spam!", Communication of the ACM, vol. 41, issue 8, Aug 1998.

[4] I. Androutsopoulos, J. Koutsias, K.V. Chandrinos, G. Paliouras, C.D. Spyropoulos, Learning to filter spam E-mail : a comparison of a naïve bayesian and a memory based approach, in: Proceedings of 4th European Conference on Principles and Practice of Knowledge Discovery in Databases, Lyon, France, September 2000, 2000, pp. 1–12.

[5] I. Androutsopoulos, J. Koutsias, K.V. Chandrinos, G. Paliouras, C.D. Spyropoulos, "An evaluation of naive bayesian anti-spam filtering", in: Proceedings of 11th European Conference on Machine Learning (ECML 2000), Barcelona, 2000, pp. 9–17.

[6] G. Sakkis, I. Androutsopoulos, G. Paliouras, V. Karkaletsis, "A memory based approach to anti-spam filtering for mailing lists", in: Empirical Methods in Natural Language Processing, 2001, pp. 44–50.

[7] G. Sakkis, I. Androutsopoulos, G. Paliouras, V. Karkaletsis, "Stacking classifiers for anti-spam filtering of E-mail", in: Kluwer academic publishers, 2003, pp. 49–73.

[8] I. Androutsopoulos, G. Paliouras, E. Michelakis, "Learning to Filter Unsolicited Commercial E-Mail", National Centre for Scientific Research Demokritos, Athens, Greece, Oct 2006.

[9] L. Zhang, J. Zhu, T. Yao, "An evaluation of statistical spam filtering techniques spam filtering as text categorization", ACM Trans. Asian Lang. Inf. Process 3 (4) (2004) 243–269.

[10] I. Koprinska, J. Poon, J. Clark, J. Chan, "Learning to classify e-mail", Inf. Sci. 177 (10) (2007) 2167–2187.

[11] G.V. Cormack, T.R. Lynam, "On-line supervised spam filter evaluation", ACM Trans. Inf. Syst. 25 (3) (2007).

[12] T.A. Almeida, A. Yamakami, "Content-based spam filtering," in: The 2010 International Joint Conference on Neural Networks (IJCNN), Barcelona, 2010, pp. 1–7.

[13] T.A. Almeida, A. Yamakami, "Spam filtering:how the dimensionality reduction affects the accuracy of Naïve Bayes classifiers," in: J Internet Serv Appl, 2011, pp. 183–200.

[14] A. Attar, R.M. Rad, R.E. Atani, "A survey of image spamming and filtering techniques", Artif. Intell. Rev. 40 (1) (2011) 71–105.

[15] D. DeBarr, H. Wechsler, "Spam detection using random forest", in: Pattern recognition letters, Elsevier, 2012, pp-1237-1244.

[16] B. Issac, "Spam detection approaches with case study implementation on spam corpora", in: Researchgate, 2010.

[17] J.R. Mendez, F. Díaz, E.L. Iglesias, J.M. Corchado, "A comparative performance study of feature selection methods for the anti-spam filtering domain", in: Advances in Data Mining. Applications in Medicine, Web Mining, Marketing, Image and Signal Mining, Springer Berlin Heidelberg, 2006, pp. 106–120.