EasyChair Preprint

№ 6074

Evaluating the Effectiveness of Machine Learning Methods for Spam Detection

Yuliya Kontsewaya, Evgeniy Antonov and Alexey Artamonov

July 13, 2021

Evaluating the Effectiveness of Machine Learning Methods for Spam Detection

Yuliya Kontsewaya[a]*, Evgeniy Antonov[a,b], Alexey Artamonov[b]

[a]*Plekhanov Russian University of Economics, Stremyanny lane 36, Moscow, 117997, Russian Federation*
[b]*National Research Nuclear University MEPhI, Kashirskoe hwy, 31, Moscow, 115409, Russian Federation*

Abstract

Technological advances are accelerating the dissemination of information. Today, millions of devices and their users are connected to the Internet, allowing businesses to interact with consumers regardless of geography. People all over the world send and receive emails every day. Email is an effective, simple, fast, and cheap way to communicate. It can be divided into two types of emails: spam and ham. More than half of the letters received by the user – spam. To use Email efficiently without the threat of losing personal information, you should develop a spam filtering system. The aim of this work is to reduce the amount of spam using a classifier to detect it. The most accurate spam classification can be achieved using machine learning methods. A natural language processing approach was chosen to analyze the text of an email in order to detect spam. For comparison, the following machine learning algorithms were selected: Naive Bayes, K-Nearest Neighbors, SVM, Logistic regression, Decision tree, Random forest. Training took place on a ready-made dataset. Logistic regression and NB give the highest level of accuracy – up to 99%. The results can be used to create a more intelligent spam detection classifier by combining algorithms or filtering methods.

Peer-review under responsibility of the scientific committee of the 2020 Annual International Conference on Brain-Inspired Cognitive Architectures for Artificial Intelligence: Eleventh Annual Meeting of the BICA Society
Keywords: Spam, Spam Filtering Method, Naive Bayes, K-Nearest Neighbors, SVM, Logistic regression, Decision tree, Random forest;

1. Introduction

Spam is one of the main problems on the Internet. The amount of spam has increased significantly over the past few years. Today, more than 85% of mails or messages received by users are spam [1]. Manual analysis of spam messages is impractical due to the large size of the data. The most accurate spam classification can be achieved using machine learning methods. Spam email called as junk email or unsolicited message which sent by spammer through Email [2]. The main types of unsolicited emails:

- Advertising;
- Nigerian spam - spam used by fraudsters to extort money from the recipient of a letter;
- Phishing is an act that attempts to electronically obtain delicate or confidential information from users (usually for the purpose of theft) by creating a replica website of a legitimate organization [3]. This message does not contain a real return address.

* Corresponding author.
E-mail address: Koncevayayul.Yu@edu.rea.ru

1877-0509

The last two types of mails are the most dangerous, as they can serve as ways to spread viruses or malware. Spam can be used as a tool for obtaining user privileges in the system, with subsequent infiltration into this system and carrying out an attack. The main way to influence the user is manipulation. The user's attention is focused on the immediate execution of actions. The main methods of luring: money, promised payments or even interests of people.

During the pandemic caused by Covid-19, users were encountering spam problem more often. One could come across the following subject of letters: tourist phishing - using fake sites to lure users, transaction emails, Facebook grants - under the Facebook logo, users were attracted with compensation funds for those who suffered during the pandemic. Russia was ranked first by the amount of outgoing spam - 23,52% [4].

Research on this topic was carried out on different dataset. Mohammed et al. used Email - 1431 dataset and applied algorithms Naive Bayes, SVM, KNN, Decision Tree. The following accuracy results were obtained 85,96%, 82,96%, 77,04%, respectively [5]. DeBarr et al. practiced Custom Collection dataset, implemented Random Forest algorithm and achieved 95.2% accuracy [6]. Nevertheless, the problem of spam remains relevant today, as spammers began to use several tricky methods to overcome the filtering methods like using random sender addresses [7] or new words.

In this paper algorithms of machine learning will be considered as modern method of spam filtering.

2. Spam Filtering Methods.

The following spam filtering methods are the most popular nowadays:

- *Systems with a confirmation request.* The sender is prompted to take some action to ensure delivery of the original message, otherwise the message is considered undelivered.
- *Use of temporary mailing addresses.* The user changes the address in case of a large number of incoming letters.
- *Blacklist*. When an incoming message arrives, the spam filter checks to see if it's IP or email address is on the blacklist; if so, the message is considered spam and rejected [8].
- *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.
- *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.
- *Linguistic heuristics*. Search in the body of the message for keywords and phrases that allow attributing this message to spam.

3. Machine Learning Approach to E-mail Spam filtering

Machine learning is an extensive subset of AI that studies how to build algorithms that can learn and make predictions, in other words, extracting patterns from examples. In the next section we will take a look at some of the most popular machine learning techniques for classifying spam, as well as existing approaches to spam detection.

3.1. Machine learning stages

To get started we will describe the main stages of machine learning process. Firstly, analysis stage: at this stage, we work with data that are processed, analyzed, and patterns are revealed. Secondly, train stage: machine learning models are used on the obtained data. The quality of the models can be improved through the selection of hyperparameters. The following stage is testing: Machine learning models are tested on unused data. Various metrics can be used to evaluate the model. The last stage is application: deployment of the best model.

3.2. The Algorithms

There are six widely spread classification algorithms in machine learning which were selected: Naive Bayes, K-Nearest Neighbours, Support Vector Machine, Logistic regression, Decision tree, Random forest.

3.2.1. Naive Bayes (NB)

Naive Bayes is a probabilistic algorithm that does a good job of classifying spam. It is called 'naive' because it ignores possible dependencies or correlations among inputs and reduces a multivariate problem to a group of univariate problems [9]. The disadvantage of this algorithm for working with spam emails is the following: the letter may contain a word that has never been found in the training sample, which will have a negative effect on the quality of classification.

3.2.2. K-Nearest Neighbours (KNN)

KNN is a metric classification method, that is, objects are represented as points in space and distances are calculated between them. Then, it enters a learning phase when training data points are iteratively assigned to a cluster which center is located at the nearest distance [10]. The number k is an input parameter of the algorithm and can be optimally tuned. The classification accuracy depends on the chosen k value. For the classified object, there are k of its nearest neighbours in the training sample. An object belongs to the class that is most common among its nearest k neighbours. The algorithm is unstable to outliers and not working correctly with a large amount of features.

3.2.3. Support Vector Machine (SVM)

SVM is a linear classifier that is equivalent to finding the hyperplane separating the classes with maximum indentation. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on [10]. The classifier tries to increase the distance between the points for the greatest "confidence" in the class definition. The model stands out for sustainability to the outliers.

3.2.4. Logistic regression

Logistic regression is a proper analysis method to model the data and explain the relationship between the binary response variable and explanatory variables. The result is the probability of assigning a value to a certain class, which is limited to values between 0 and 1 [11].

3.2.5. Decision tree (DS)

A decision tree is a hierarchical organized structure that divides the feature space into a subspace, where a prediction result is issued for each object in this subspace. Algorithm tends to overfit.

3.2.6. Random forest (RF)

Random forest is a prediction algorithm that uses the idea of constructing trees. A forest, that is, several trees, enhance the predictive ability of each tree separately. During the training phase, a number of decision trees are constructed (as defined by the programmer) which are then used for the class prediction; this is achieved by considering the voted classes of all the individual trees and the class with the highest vote is considered to be the output [3].

3.3. Approaches and Experiment Description

There are two main approaches to spam detection:

1) Classification of images or other attachments using machine learning algorithms to identify threats.
2) Natural language processing to analyze the text of an email in order to detect spam.

The second approach is chosen to build the spam detection algorithm. Fig. 1 describes the main stages of machine learning and the subpoints to which the data is exposed.

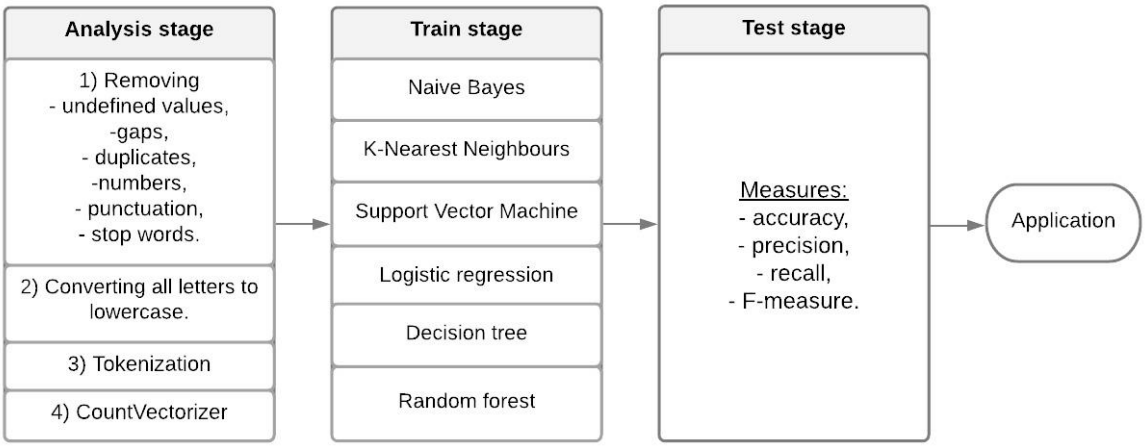

Fig. 1. Scheme of the Developed Algorithm

In the first stage of machine learning, the data will be cleaned, then separated a sentence into words and built a system of attributes for a text. After that, the selected algorithms will be trained and their accuracy will be assessed using measures such as accuracy, precision, recall and F-measure. In this way, an algorithm will be selected that better solves the problem of spam classification.

4. Result

The training took place on a ready-made labeled dataset, where class 1 indicates that the message is spam, and class 0 that is does not. This data is similar to the data collected on the mail server. The main stages of creating a classifier are:

1) Data collection;
2) Pre-processing and text cleaning;
3) Learning and obtaining prediction results.

The *sklearn* library will be used to work with machine learning algorithms. The main task of the algorithm is to find patterns.

In order to assess the quality of the selected algorithms, a ready-made dataset from the site *kaggle.com* [13] was selected. The file *emails.csv* is a set of messages in English in the amount of 5728 emails, each of which is spam or ham (Fig. 2). The dataset consists of two columns such as the message text and the result of classifying the message as spam or ham, where 1 indicates that the message is spam and 0 - ham.

	text	spam
0	Subject: naturally irresistible your corporate...	1
1	Subject: the stock trading gunslinger fanny i...	1
2	Subject: unbelievable new homes made easy im ...	1
3	Subject: 4 color printing special request add...	1
4	Subject: do not have money , get software cds ...	1
...	...	...
5723	Subject: re : research and development charges...	0
5724	Subject: re : receipts from visit jim , than...	0
5725	Subject: re : enron case study update wow ! a...	0
5726	Subject: re : interest david , please , call...	0
5727	Subject: news : aurora 5 . 2 update aurora ve...	0

5728 rows × 2 columns

Fig. 2. Dataset

Supervised learning is a method in which a pair (object, response) is fed to the model input. It is required to find the dependence of response on objects.

4.1. Data processing and algorithm creation process

The first stage of data processing and algorithm creation process is to prepare data. First of all, there is a need to remove undefined values, gaps and duplicates in the data. Despite the fact that many models allow the use of undefined data and gaps in the sets, it is better to remove them before performing basic manipulations to eliminate possible errors and improve the quality of the classification.

Let's look at the distribution of spam and non-spam using the seaborn library (a library for generating statistical graphics in Python). The data is a collection of 4360 non-spam samples and 1368 spam messages (see Fig. 3). After removing duplicates, new dataset size is: (5695, 2).

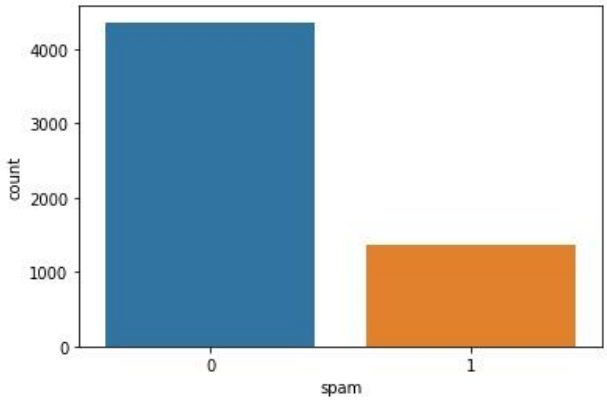

Fig. 3. Quantitative distribution of spam and non-spam messages

The next stage is to check for missing values. No gaps were found in the data, we can proceed to the next stage, which is text processing.

Text processing is an important step in machine learning, as data can contain a lot of noise and unwanted characters such as punctuations, numbers. To clear the data, the following steps were followed:

- convert all letters to lowercase,
- delete numbers includes,
- remove punctuation marks and stop words (useless words - prepositions, pronouns, etc.).

Punctuation and stopwords have been removed with the help of the *string* library and *nltk*.

The fourth stage of the process is a tokenization. Tokenization means splitting a sentence into words that are separated by a comma.

Algorithms always "expect" the input to be integers/floating point numbers, therefore, further it is necessary to extract features. The CountVectorizer method was used to convert words to numbers. CountVectorizer is a function that builds a system of attributes for a text. For each word from the dictionary, a binary sign is set up. That is, the text is converted into a matrix consisting of the number of tokens.

Further, we separate the data into test and training sets. The data set was divided into training and test samples to control the operation of the algorithm, 80% and 20%, respectively.

In the last stage of the process, we assess the quality of the classification. Model training is the process of running the selected algorithm based on the training data (with known message prediction values: spam/ham). The quality of the algorithms that give out belonging to classes will be assessed using the following measures: precision, recall, accuracy and F-measure. In the case of choosing one of two classifiers (for example, if the precision of the first classifier is greater than the second, but the recall of the first is less), we can use the F-value, which correctly aggregates the accuracy and completeness into one expression.

For each classification algorithm, except for the Naive Bayesian classifier, hyperparameter optimization was performed using GridSearchCV. The considered parameters and values are summarized in Table 1.

Table 1. Hyperparameters and their possible values

Algorithm	Hyperparameters and their possible values
KNN	'n_neighbors': (5,6,7,8,9,10)
Naive Bayes	-
Decision tree	'max_depth': np.arange(1, 10),'criterion': ['gini','entropy']
SVM	'C':[1,10,100,1000], 'gamma':[1,0.1,0.001,0.0001], 'kernel':['linear','rbf']
Random forest	'n_estimators': [10, 20, 30], 'max_depth': [2, 5, 7, 10]
Logistic regression	“C”:logspace(-3, 3, 7), “penalty”:[“l1”, “l2”]}

In the experiment, the following results were obtained and entered into Table 2. The measures NB and Logistic regression are higher compared to KNN, Decision tree, SVM, Random forest.

Table 2. Results of the algorithms

Algorithm	Accuracy	Precision	Recall	F-measure
KNN	0,90	0,91	0,63	0,74
Naive Bayes	0,99	0,97	0,99	0,98
Decision tree	0,94	0,82	0,96	0,88
SVM	0,98	0,98	0,95	0,96
Random forest	0,84	1	0,28	0,42
Logistic regression	0,99	0,98	0,96	0,97

5. Conclusion

To sum up, we can conclude that spam is a disturbing problem which is being faced by every user, that's why we have performed research on this topic. There are several ways to build spam protection. Machine learning algorithms provide high accuracy in the classification of spam messages. In this research training took place on a ready-made dataset. The data is a collection of 4360 non-spam samples and 1368 spam messages.

These results were achieved throughout machine learning models such as: Naive Bayes, K-Nearest Neighbors, SVM, Logistic regression, Decision tree, Random forest. In this paper we have demonstrated that for spam filtering the most efficient algorithms are Logistic regression and NB give as they have the highest level of accuracy. It has reached 99%. The results can be used to create a more intelligent spam detection classifier by combining algorithms or filtering methods.

References

[1] M. Bassiouni, M. Ali & E. A. El-Dahshan (2018) "Ham and Spam E-Mails Classification Using Machine Learning Techniques." *Journal of Applied Security Research* **13(3)**: 315-331.

[2] Hanif Bhuiyan, Akm Ashiquzzaman, Tamanna Islam Juthi, Suzit Biswas and Jinat Ara (2018) "A Survey of Existing E-Mail Spam Filtering Methods Considering Machine Learning Techniques.", *Global Journal of Computer Science and Technology* **18(2)**: 20-29.

[3] Andronicus A. Akinyelu1 and Aderemi O. Adewumi (2014) "Classification of Phishing Email Using Random Forest Machine Learning Technique." *Journal of Applied Mathematics* **2014**:1-6.

[4] Tatyana Kulikova & Tatyana Sidorina (2020) "Spam and phishing in Q3 2020", in Security list by Kaspersky https://securelist.com/spam-and-phishing-in-q3-2020/99325/, last accessed 2021/01/10

[5] Mohammed, S., Mohammed, O., Fiaidhi, J., Fong, S. J., & Kim, T. H. (2013) "Classifying Unsolicited Bulk Email (UBE) using Python Machine Learning Techniques." *International Journal of Hybrid Information Technology* **6(1)**: 43-56.

[6] DeBarr, D. & Wechsler, H. (2009) "Spam detection using clustering, random forests, and active learning.", in Sixth Conference on Email and Anti-Spam. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.537.1694&rep=rep1&type=pdf, last accessed 2021/01/12

[7] Cormack, Gordon. Smucker, Mark. Clarke, Charles (2011) "Efficient and effective spam filtering and re-ranking for large web datasets." *Information Retrieval* **14**: 441–465.

[8] Akshay Prakashrao Gulhane (2013) "Spam Filtering Methods for Email Filtering." *International Journal Of Computer Science And Applications* **6(2)**: 18-23.

[9] Anurag Sinha & Shubham Singh (2020) "A Detailed study on email spam filtering techniques.", *International Journal of Data Science and Analytic* **10(3)**:1-34.

[10] Hedieh Sajedi, Golazin Zarghami Parast, Fatemeh Akbari (2016) "SMS Spam Filtering Using Machine Learning Techniques: A Survey." *Machine Learning Research* **1(1)**:1-14.

[11] Gaurav Chauhan (2018) "All about Logistic regression in one article", in towards data science https://towardsdatascience.com/logistic-regression-b0af09cdb8ad, last accessed 2021/01/11

[12] Prashant Gupta (2017) "Decision Trees in Machine Learning", in towards data science https://towardsdatascience.com/decision-trees-in-machine-learning-641b9c4e8052, last accessed 2021/01/11

[13] Spam filter:Identifying spam using emails, https://www.kaggle.com/karthickveerakumar/spam-filter, last accessed 2021/01/11