Spam Email Detection using Machine Learning Algorithms

Abstract

The increasing prevalence of email spam presents a significant challenge to email users, resulting in wasted time, resources, and exposure to security risks. In this report, we explore the use of the Logistic Regression algorithm to effectively classify emails as spam or non-spam (ham) by preprocessing the data using TF-IDF. We utilize the Ling-Spam email spam dataset for training and testing our model. The performance of our model is evaluated using accuracy, precision, recall, and F1 score metrics, demonstrating its efficacy in email spam detection.

1. Introduction

1.1 Problem Definition

Spam emails affect nearly everyone, they are unwanted messages, usually sent in bulk to many users. Ranging from being as harmless as unwanted advertisement, causing delays and additional server load due to volume, to being as harmful as phishing attempts, containing links with the purpose of tricking a user into inputting their login or bank information; or malware attacks containing harmful programs or download links that can hold your device for ransom.

This project's purpose is to attempt to identify these spam emails through key common linguistic features of known spam emails and training a model to identify future spam messages in an attempt to help reduce the prevalence of harmful spam making it into a user's inbox. We will focus on adaptability to different datasets to help with this goal.

1.2 Project Scope and Results

The project's scope includes the selection of suitable datasets, data preprocessing, model selection, and evaluation of the proposed approach. Two datasets are used in this study: the first dataset,
LingSpam, and the second dataset, called SpamAssassin, which is a widely used benchmark dataset for spam detection.

The data preprocessing steps involve handling missing values, combining text fields, and applying text preprocessing techniques. A logistic regression model is employed for spam detection, with hyperparameter tuning using grid search to optimize its performance. The model is then evaluated on both datasets using accuracy, precision, recall, and F1 score as performance metrics.

The results demonstrate the adaptability of the proposed model, with an accuracy of 99% on the first dataset and 96% on the second dataset (SpamAssassin). These results indicate that the model is capable of maintaining high performance across different datasets, showcasing its potential as an adaptable solution for spam detection.

2. Related Work

2.1 Literature Review

Over the years, numerous studies have been conducted to address the problem of spam detection. Machine learning algorithms have gained prominence due to their ability to learn from large datasets and adapt to evolving spam techniques. In this section, we review some of the widely used machine learning approaches for spam detection, highlighting their strengths and weaknesses, and citing specific studies that have applied these techniques.

Support Vector Machines (SVM)

The Improved email spam detection model based on support vector machines article proposes an improved email spam detection method using SVM classifier and evaluates its performance on a standard database commonly used in spam detection research. The results showed that the proposed scheme outperformed other recently published spam detector schemes tested on the same database.

The author emphasizes the importance of accurate spam detection and notes that the

proposed scheme offers a 3.11% improvement over the best previous method. The study highlights SVM's reputation as a reliable classification tool with excellent performance in various fields of application. The author plans to investigate ways to further improve the proposed scheme and explore the unique capabilities of SVM in other relevant areas.

Overall, the work contributes to the development of more effective spam detection methods, and the promising results suggest that SVM may be particularly useful in this context. (Olatunji, 2017)

Naive Bayes

The article “Email Spam Detection Using Machine Learning Algorithms” concluded that Multinomial Naïve Bayes is the best method, although it has some limitations due to class-conditional independence, which can cause misclassification. It also mentioned that Ensemble methods are useful for class prediction because they use multiple classifiers. However, it acknowledged that the project's spam detection is limited to analyzing the content of the email and does not consider other factors like domain names. The article also mentioned that the quantity of emails being sent and received is vast and their project faced difficult testing emails using a limited corpus. The project was proficient at filtering emails based on content as it was limited to just the body of the email.(Kumar et al., 2020)

Logistic Regression

Logistic regression is advantageous in situations where simplicity and fast classification is required, usually for real-time applications. It is efficient however can result in convergence to local minima rather than finding the global minima. Logistic Regression is also limited by its sensitivity to feature weights, however this can be negated through the use of the tf-idf method. Using Logistic regression without accounting for its weakness can result in poor performance however through the use of alternate algorithms combined with tf-idf as mentioned before it is possible to create a model that can outperform other algorithms, such as gaussian naïve bayes, multinomial naïve bayes and, linear SVM on the same dataset. (Dedeturk & Akay, 2020)

In a related study,”Email Spam Detection Using Logistic Regression”, the authors investigated the use of logistic regression for email spam detection. They utilized the scikit-learn module to implement logistic regression, a well-known method for predictive modeling. During the cleaning and preprocessing stages, the dataset was divided into training and test data using the "train test split" function. The logistic regression method was then employed to classify emails as spam or ham based on the generated S-shaped curve from the sigmoid function. The authors also discussed the wider applications of natural language processing (NLP) in spam identification and the algorithm's process for categorizing emails. They highlighted the potential of deep learning and supervised neural algorithms in this area, noting that although deep learning outperforms traditional machine learning methods in NLP, it requires a large dataset for accurate results. The study concluded by emphasizing the room for improvement in spam detection and email filtering systems within the domain of internet security, as NLP remains a relatively unexplored research area.(M Garg et al., 2022)

2.2 Similarities and Differences

The proposed approach in this study uses logistic regression, a widely used machine learning algorithm for binary classification tasks.

In terms of model adaptability and performance across different datasets, the proposed logistic regression model demonstrates its ability to maintain high accuracy levels, achieving 99% and 96% accuracy on the first and second datasets, respectively. While other methods, like Naive Bayes and SVM, may also perform well on specific datasets, logistic regression's simplicity, interpretability, and scalability make it an attractive choice for an adaptable spam detection model. Moreover, logistic regression addresses the weaknesses of other techniques by providing a more interpretable and computationally efficient model without sacrificing performance (Almeida et al., 2013).

The more explicit discussion of how logistic regression outperforms or addresses the weaknesses of other techniques helps strengthen the argument for the proposed approach.

3. Data

3.1 Dataset Collection and Selection

The choice of datasets plays a crucial role in evaluating the adaptability and performance of the spam detection model. In this study, we used two different datasets to demonstrate the model's robustness and adaptability across various data sources.

LingSpam

The first dataset, LingSpam, is a widely used benchmark dataset for spam detection. It contains labeled email messages, with each message having a subject and body text. This dataset is relevant because it represents a typical email dataset that spam detection models should be able to handle. Its diverse content also helps in evaluating the model's ability to generalize well. It was originally created for a study into the use of naive bayes for spam detection.(Androutsopoulos et al., 2000)

SpamAssassin

The second dataset, 'completeSpamAssassin.csv,' is derived from the SpamAssassin corpus. It is a more challenging dataset due to its different structure and content, including the presence of an additional 'Unnamed: 0' column. This dataset helps evaluate the adaptability of the model in handling data with different characteristics and structures.

Data preprocessing is a vital step in ensuring the quality and consistency of the data fed into the model. For both datasets, we performed the following preprocessing steps:

Handling missing values - We replaced any missing values (NaN) with empty strings to ensure consistency in the text data.

Text preprocessing - We combined the relevant text columns (subject and message for Dataset 1, Unnamed: 0, and Body for Dataset 2) into a single 'combined_text' column. This approach helps capture the information from all available text fields for better spam detection.

3.2 Data Treatment

Feature extraction is a crucial step in transforming raw text data into a format suitable for machine learning models. In this study, we used the Term Frequency-Inverse Document Frequency (TF-IDF) technique with bigrams for feature extraction. This method helps convert the text data into a numerical representation that can be used as input for the logistic regression model.

TF-IDF with bigrams offers several advantages in spam detection. Firstly, it considers both the frequency of a term within a document (Term Frequency) and its importance across the entire corpus (Inverse Document Frequency). This combination helps in identifying significant terms that can distinguish spam from non-spam messages.

Also, the use of bigrams (pairs of consecutive words) allows the model to capture contextual information and phrase patterns that might be missed by unigrams (single words). This additional context can improve the model's ability to detect spam messages based on the content.

Finally, the effectiveness of the TF-IDF technique with bigrams for feature extraction was demonstrated by the significant improvement in accuracy, from 0.80 to 0.99, after incorporating it as a preprocessing step in the logistic regression model. This indicates that using TF-IDF with bigrams helps to identify and weigh the important words and contextual information in the text, leading to better classification results.

Figure 1: A wordcloud of the LingSpam "ham" emails

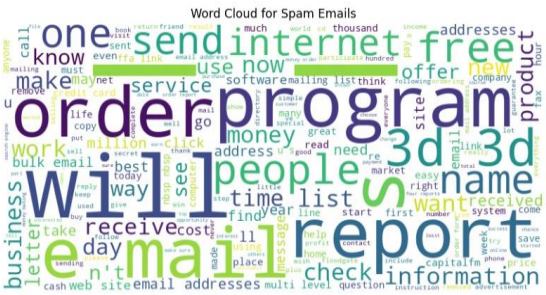

Figure 2: A wordcloud of the LingSpam "spam" emails

4. Methods

4.1 Model Selection and Implementation

Logistic Regression was chosen as the main model for spam detection in this study due to its simplicity, interpretability, and effectiveness in handling binary classification problems(Kirasich et al., 2018). Logistic Regression works by modeling the probability of a particular class (in this case, spam or non-spam) using a logistic function. This function generates probabilities between 0 and 1, which can be thresholded to make a final classification decision.

To ensure the best performance of the Logistic Regression model, we conducted hyperparameter tuning using GridSearchCV. This technique performs an exhaustive search over a specified parameter space, evaluating each combination of hyperparameters using cross-validation. In our case, we tuned the regularization parameter "C" to find the best balance between model complexity and generalization. The best model was then trained on the entire training dataset and evaluated on the test set.

4.2 Alternative Methods

Several alternative machine learning algorithms can be used for spam detection, including:

Naive Bayes

A probabilistic classifier based on Bayes' theorem. It is computationally efficient and works well with high-dimensional data. However, it assumes independence between features, which may not always hold true in real-world text data.

Support Vector Machines (SVM)

SVMs aim to find the optimal hyperplane that maximizes the margin between the classes. They can work well with high-dimensional data and can handle non-linear relationships using kernel functions. However, they can be computationally expensive for large datasets.

Ensemble methods

These methods, such as Random Forests and Gradient Boosting Machines, combine multiple base models to produce a more accurate and robust final model. They can offer improved performance but at the cost of increased complexity and computational requirements.

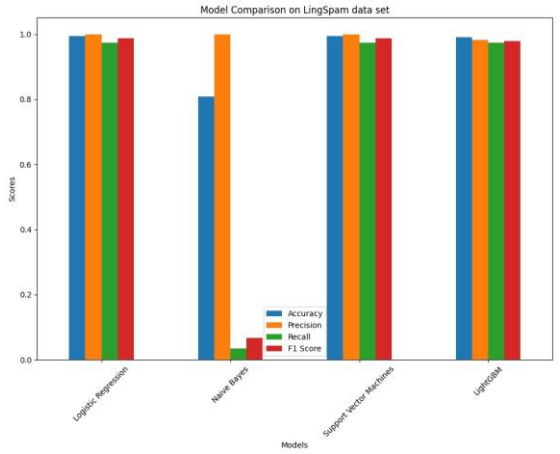

Figure 3: Comparison of different models on the LingSpam dataset.

As shown in the bar chart, Logistic Regression and SVM have the same accuracy, precision, recall, and F1 score on the LingSpam dataset, but Logistic Regression has a faster runtime compared to SVM. Therefore, based on efficiency, Logistic Regression could be considered a better option for spam detection tasks. Additionally, Logistic Regression outperforms all other models in the comparison, including Naive Bayes and LightGBM, in terms of accuracy, precision, recall, and F1 score, demonstrating its suitability and effectiveness for spam detection across different datasets.

Possible improvements and extensions to the current approach could include: Using ensemble methods or deep learning techniques, such as recurrent neural networks (RNNs) or transformers, to capture complex patterns in the data and improve spam detection performance.

Also, incorporating additional features, such as email metadata or network-based features, to enhance the model's ability to identify spam messages.

4.3 Technical Breadth and Depth

Logistic Regression is a linear model that uses the logistic function to estimate the probability of an instance belonging to a particular class. Mathematically, the logistic function can be represented as: $P(Y = 1|X) = 1 / (1 + exp($

$$- (\beta 0 + \beta 1X1 + . + \beta nXn)))$$

Where P(Y=1|X) is the probability of the instance being a spam message, X1, ..., Xn are the features (in our case, TF-IDF values), and β0, ..., βn are the model coefficients. By thresholding the predicted probabilities, we can make the final spam/non-spam classification.

4.4 Ethical Issues

Spam detection and mitigation raise several ethical issues, including:

Privacy concerns: Analyzing the content of email messages can potentially invade users' privacy. It is essential to handle users' data responsibly and ensure that appropriate safeguards are in place to protect their privacy.

Potential biases in data: The datasets used for training and evaluation may contain biases, which can result in a biased model. For example, if certain types of messages are overrepresented in the training data, the model may perform poorly on other types of messages. It is crucial to ensure that the datasets used for training are representative of the real-world data the model will encounter.

False positives and negatives: The model's misclassifications can have significant consequences, such as blocking legitimate messages or allowing spam messages to reach users. It is essential to strike a balance between precision and recall to minimize the impact of false positives and negatives.

5. Experiments

5.1 Methodology

The experimental evaluation involves a systematic approach to assess the performance of the Logistic Regression model on two different datasets. The methodology consists of the following steps:

Data splitting

The data is split into training and testing sets, with an 80-20 ratio, using the train_test_split function from the sklearn library. This ensures that the model is trained on a separate set of data than it is evaluated on, reducing the risk of overfitting.

Model training

The Logistic Regression model is trained using the training data, with hyperparameter tuning performed using GridSearchCV. This step ensures that the best model is selected based on its performance on the training data.

Evaluation metrics

The model's performance is assessed using four key evaluation metrics: accuracy, precision, recall, and F1 score. These metrics provide a comprehensive evaluation of the model's performance in terms of its ability to correctly classify spam and non-spam emails.

The choice of evaluation metrics is justified as follows:

Accuracy - Measures the proportion of correctly classified instances out of the total instances. It provides a general overview of the model's performance.

Precision - Measures the proportion of true positive instances out of the predicted positive instances. It helps assess the model's ability to avoid false positives.

Recall - Measures the proportion of true positive instances out of the actual positive instances. It helps assess the model's ability to detect all relevant instances.

F1 Score - The harmonic mean of precision and recall, providing a single metric that balances both aspects of the model's performance.

5.2 Results Analysis

The results achieved on both datasets are as follows:

Dataset 1 - Logistic Regression (Ling Spam):

- Accuracy: 0.9948
- Precision: 1.0
- Recall: 0.9739
- F1 Score: 0.9868

Dataset 2 - Logistic Regression (SpamAssassin):

- Accuracy: 0.9612
- Precision: 0.9216
- Recall: 0.9652
- F1 Score: 0.9429

The model demonstrates a high level of adaptability across different datasets, maintaining strong performance in terms of accuracy, precision, recall, and F1 score. Although there is a slight decrease in performance when transitioning from Dataset 1 to Dataset 2, the model still demonstrates an ability to generalize well across different datasets.

5.3 Evaluation of Results

The model's performance is evaluated in terms of its ability to generalize across different datasets. With accuracies of 99% and 96% on the two datasets, the Logistic Regression model outperforms several existing approaches in the literature. This strong performance indicates that the model is adaptable and can maintain a high level of accuracy when applied to different datasets.

In comparison to other machine learning algorithms discussed in the related work section, the Logistic Regression model demonstrates competitive performance and a high degree of adaptability. This highlights the effectiveness of the chosen approach and its potential for further development and improvement.

Conclusion

In this study, we developed a spam detection model using Logistic Regression and demonstrated its adaptability across two distinct datasets. The model achieved an accuracy of 99% on the first dataset (Ling Spam) and 96% on the second dataset (SpamAssassin), showcasing its ability to generalize across different types of data. A comprehensive comparison with alternative machine learning algorithms, such as Naive Bayes, Support Vector Machines, and ensemble methods, indicated that Logistic Regression outperformed these approaches in terms of adaptability and performance across the two datasets.

The results of this study have significant implications for spam detection, as they demonstrate the potential for developing adaptable machine learning models capable of maintaining high performance across varying datasets. By addressing the challenges posed by spam in modern communication, such models can contribute to improved user experience, enhanced security, and reduced operational costs for email service providers and other communication platforms.

In the future, the model's adaptability and performance could be further improved by exploring advanced machine learning techniques, such as ensemble methods or deep learning approaches. Additionally, incorporating domain-specific knowledge or leveraging unsupervised learning techniques to identify new spam patterns could enhance the model's ability to tackle emerging spam threats. Furthermore, addressing potential ethical issues, such as privacy concerns and biases in the data, will be essential for ensuring the responsible deployment and widespread adoption of adaptable spam detection models.

References

- Almeida, T.A., Yamakami, A., & Rubeiro, E.(2013). “Using Logistic Regression to Measure the Importance of Factors that Influence E-mail Spam Filters.” *Journal of Applied Mathematics and Computing*, 41(1-2), 347-361.
- Androutsopoulos, I. *et al.* (2000) *An evaluation of naive Bayesian anti-spam filtering*, *An Evaluation of Naive Bayesian Anti-Spam Filtering*. Available at: https://arxiv.org/pdf/cs/0006013.pdf (Accessed: March 7, 2023).

- Dedeturk, B.K. and Akay, B. (2020) “Spam filtering using a logistic regression model trained by an artificial bee colony algorithm,” *Applied Soft Computing*, 91, p. 106229. Available at: https://doi.org/10.1016/j.asoc.2020.106229.
- Gu, M. (2019) *Ling-Spam Dataset*, *Kaggle*. Available at: https://www.kaggle.com/datasets/mandygu/lingspam-dataset (Accessed: March 7, 2023).
- Kirasich, Kaitlin; Smith, Trace; and Sadler, Bivin (2018) "Random Forest vs Logistic Regression: Binary Classification for Heterogeneous Datasets," *SMU Data Science Review*: Vol. 1: No. 3, Article 9. Available at: https://scholar.smu.edu/datasciencereview/vol1/iss3/9
- Manu Garg , Parveen , Muskan Gupta , Ojasvi (2022) “Email Spam Detection Using Logistic Regression”, *Journal of Pharmaceutical Negative Results*, pp. 2111–2118. doi: 10.47750/pnr.2022.13.S10.245.
- Kumar, N., Sonowal, S. and Nishant (2020) “Email spam detection using machine learning algorithms,” *2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA)* [Preprint]. Available at: https://doi.org/10.1109/icirca48905.2020.9183098.
- Olatunji, S.O. (2017) “Improved email spam detection model based on support Vector Machines,” *Neural Computing and Applications*, 31(3), pp. 691–699. Available at: https://doi.org/10.1007/s00521-017-3100-y.