Email Spam Detection

AI IN CYBER SECURITY

23-December-2025

Abstract

Spam emails are unsolicited messages that pose a threat to users by wasting time and potentially delivering malicious content. This study compares the performance of two machine learning algorithms, Naive Bayes and Logistic Regression, for detecting spam messages using the SMS Spam Collection Dataset. The dataset consists of 5,574 labelled messages classified as spam or ham (non-spam). Text data was preprocessed using TF-IDF vectorisation, and the models were evaluated using accuracy, precision, recall, and F1-score. The results show that both algorithms performed effectively in classifying messages; however, Naive Bayes achieved higher overall performance across most evaluation metrics. Logistic Regression also demonstrated reliable classification performance, particularly in precision. These findings suggest that both algorithms are suitable for spam detection tasks, with Naive Bayes offering stronger overall performance for this dataset.

Introduction

Spam email is unsolicited and unwanted junk email sent out in bulk to an indiscriminate recipient list. Typically, spam is sent for commercial purposes. It can be sent in massive volume by botnets, networks of infected computers.

Sometimes, these emails are sent for the purpose of advertising. Though considered unethical, some companies still make use of these emails. The cost that these emails incur is extremely low, and they can send a mass amount every time. These emails can also be considered as a malicious attempt to get access to your computer.

Spam messages can be hard to prevent, especially because spam messages can also come from a botnet. Botnets refer to a network of already compromised computers. Therefore, it can be hard to trace and prevent the spammer.

If you receive an incoming message that seems to be spam for instance, you do not recognize or receive messages from the sender then mark that message spam in your email program. Do not click on any of the links or attachments that are provided in that spam message. This is because spammers may provide those links to confirm the authenticity of an email address or even have malicious webpages or executables downloaded. Anti-spam solutions must protect against various known threats other than spam, phishing, and botnet attacks that include difficult-to-detect, short-lived, low-volume electronic mail threats. Learn more about the features of Email Threat Defense.

"Spam" email can be dangerous. "Spam" email can contain links that actually give your computer a virus (see "What is Malware?" below). Do not follow links to "spam" messages. "Spam" messages usually try to trick you into believing they are very important messages that require immediate attention. Continue reading to discover some of the "basics" of "spam."

Reseach Question

This study aims to compare the performance of Naive Bayes and Logistic Regression algorithms for email spam detection.

Literature Review

Traditional Spam Filtering Methods

Before machine learning became prevalent, spam detection mainly relied on rule-based systems. These methods used manually created rules or keyword lists to identify unwanted messages. For instance, if an email included phrases like "free money" or "click here," it would be flagged as spam. While these approaches were easy to implement, they often had trouble adjusting. Spammers could easily bypass rules by slightly tweaking their wording, resulting in many false negatives or false positives.

Use of Machine Learning in Text Classification

Machine learning changed spam detection by allowing systems to learn from data automatically, instead of just using human-defined rules. Text classification algorithms examine patterns across numerous messages, recognizing subtle differences between spam and legitimate emails. Features such as word frequencies, n-grams, or even more advanced embeddings can train models that generalize well to new messages.

Previous Studies Using Naive Bayes and Logistic Regression

Two widely used machine learning algorithms for text classification are Naive Bayes and Logistic Regression.

Naive Bayes:

Naive Bayes classifiers assume that the presence of each word in a message is independent of the others. Despite this strong assumption, they often perform surprisingly well in text classification tasks.

Their main advantages are simplicity, quick training, and efficiency with high-dimensional data, such as text with thousands of unique words.

Naive Bayes is particularly effective for spam filtering because the probabilistic approach naturally manages imbalanced datasets, where spam messages may be less common than normal emails.

Logistic Regression:

Logistic Regression is a popular baseline classifier in text classification tasks. It predicts the likelihood that a message belongs to a certain class (spam or ham) using a linear combination of features.

It is appreciated for its clarity, reliability, and strong performance, often serving as a standard to compare against more complex models.

Unlike Naive Bayes, Logistic Regression does not assume feature independence, allowing it to capture interactions between words when properly adjusted.

By combining these algorithms with suitable feature extraction techniques, researchers and practitioners have built highly accurate spam detection systems that keep up with new types of spam over time.

Methodology

3.1 Dataset

For this study, the SMS Spam Collection Dataset was used. This dataset contains a collection of text messages that are labeled as either spam or ham (non-spam). It consists of approximately 5,574 messages, with a mix of spam and ham messages. The dataset provides a simple but effective way to test machine learning algorithms for text classification tasks. Using this dataset allows us to simulate a real-world scenario where messages need to be automatically classified to protect users from unwanted or malicious content.

3.2 Data Preprocessing

Before applying machine learning algorithms, the dataset underwent several preprocessing steps to ensure the data was clean and in a suitable format for modeling. First, text cleaning was performed, which included removing unnecessary characters, punctuation, and converting all text to lowercase to maintain consistency.

Next, the text messages were converted into a numerical representation using TF-IDF (Term Frequency–Inverse Document Frequency). TF-IDF is a widely used technique in natural language processing that transforms text into numerical vectors by measuring the importance of each word in the message relative to the entire dataset. This allows machine learning models to understand and process textual data.

Finally, the dataset was split into training and testing sets, using an 80/20 split. The training set was used to train the algorithms, while the testing set was used to evaluate their performance on unseen data. This approach ensures that the models are tested fairly and can generalize well to new messages.

3.3 Algorithms Used

Two machine learning algorithms were implemented and compared in this study: Naive Bayes and Logistic Regression.

Naive Bayes is a probabilistic classifier that assumes the features (in this case, words in the messages) are independent of each other. Despite its simplicity, Naive Bayes has proven to be highly effective for text classification tasks such as spam detection. Its ability to handle high-dimensional data makes it suitable for analyzing large text datasets.

Logistic Regression, on the other hand, is a linear classification algorithm that predicts the probability of a message belonging to a particular class (spam or ham). Logistic Regression works well for binary classification tasks and provides a strong baseline for performance comparison. By modeling the relationship between the input features and the probability of a message being spam, it can effectively separate the two classes.

3.4 Evaluation Metrics

The performance of both algorithms was evaluated using four key metrics: accuracy, precision, recall, and F1-score.

- Accuracy measures the proportion of correctly classified messages out of all messages.
- Precision measures how many messages predicted as spam are actually spam, which is important to minimize false alarms.
- Recall measures how many actual spam messages were correctly identified, ensuring important spam messages are not missed.
- F1-score is the harmonic mean of precision and recall and provides a balanced measure of a model's performance.

These metrics were chosen to give a comprehensive assessment of the algorithms, considering both their ability to correctly identify spam and to avoid misclassifying ham messages.

Result

Model	Accuracy	Precision	Recall	F1 Score
Naive Bayes	0.9668	1.0000	0.7533	0.8593
Logistic Regression	0.9525	0.9709	0.6667	0.7905

Table 1: Performance comparison of classification models

Result Brief

The performance of the two machine learning algorithms, Naive Bayes and Logistic Regression, was evaluated using accuracy, precision, recall, and F1-score on the SMS Spam Collection Dataset.

The results show that both algorithms performed well at classifying messages as spam or ham. However, Naive Bayes achieved slightly better overall results than Logistic Regression in this

experiment. Naive Bayes recorded higher accuracy, recall, and F1-score, meaning it was more effective at correctly identifying spam messages while maintaining reliable classification performance.

Logistic Regression also showed strong performance, particularly in precision, indicating that when it predicted a message as spam, it was usually correct. Although its overall scores were slightly lower than Naive Bayes, it still proved to be a reliable model for spam detection.

The comparison graph clearly highlights these results and demonstrates that both algorithms are suitable for spam detection tasks. Overall, the findings suggest that Naive Bayes may be more effective for this dataset, especially where fast computation and strong recall are important, while Logistic Regression remains a good alternative when interpretability and consistent prediction behaviour are preferred.

Discussion

The findings from this research show that both Naive Bayes and Logistic Regression classified spam messages effectively. However, the results indicate that Naive Bayes outperformed Logistic Regression across most evaluation metrics, including accuracy, recall, and F1-score. This suggests that Naive Bayes was more effective at identifying spam messages within the SMS Spam Collection Dataset.

One reason for this outcome may be that Naive Bayes performs particularly well on text classification tasks when combined with TF-IDF features. Although Naive Bayes assumes independence between words, this limitation did not significantly affect performance for this dataset. Logistic Regression also produced strong results, especially in precision, meaning that when it classified a message as spam, it was usually correct. However, its overall performance was slightly lower than that of Naive Bayes in this experiment.

The size and attributes of the dataset also influence the models' results. The SMS Spam Collection Dataset, which contains a moderate number of spam and ham messages, is well-suited for both models. However, a larger dataset may yield better results for Logistic Regression, which may need parameter optimization. The text features extracted from this dataset using TF-IDF were crucial for building effective models, as they highlight the importance of words relative to the entire dataset, helping to distinguish spam from ham.

The limitations of this analysis include the use of one dataset and only two algorithms. While the SMS Spam Collection Dataset is commonly used, it may not represent all types of spam messages that occur in real-world scenarios. Additionally, other algorithms like SVM or Neural Networks may produce different outcomes, and using ensemble learning could improve the results further. Lastly, although the preprocessing technique employed is standard, it did not incorporate advanced feature engineering methods that might enhance performance.

Critical Reflection

The results of this study show that both Naive Bayes and Logistic Regression work well for spam detection, but there are several limitations to consider. One significant issue is the imbalance in the dataset, with fewer spam messages than ham messages. This imbalance can lead models to favor predicting ham, which might lower recall for spam. Therefore, relying solely on accuracy can be misleading; additional metrics like precision, recall, and F1-score were needed for a fairer evaluation of performance.

Data quality is another limitation. Text messages can include spelling errors, unusual formatting, or other issues that affect TF-IDF vectorization and lower model accuracy. Additionally, the SMS Spam Collection Dataset contains around 5,500 messages. While this is enough for comparison, it may not capture the complexity of real-world spam. Consequently, the models might not generalize well to new or evolving spam patterns.

There are also concerns specific to the models. Logistic Regression may overfit when using many text features without careful regularization. Naive Bayes, though efficient, assumes that words operate independently, which oversimplifies language and may hinder performance when word interactions matter. Moreover, the difference in performance between the two models was relatively small, which might lessen the significance of the comparison.

Finally, deploying these models in the real world poses further challenges. Spam patterns evolve over time, and biased datasets may not cover all types of spam. Ethical issues must also be taken into account, as overly aggressive filtering could block legitimate messages.

Conclusion

This paper compared the performance of Naive Bayes and Logistic Regression for spam detection using the SMS Spam Collection Dataset. Both machine learning models demonstrated strong performance in classifying messages as spam or ham. However, the results showed that Naive Bayes achieved better overall performance, particularly in terms of accuracy, recall, and F1-score. Logistic Regression remained a reliable alternative, offering strong precision and consistent classification behaviour. These findings indicate that both algorithms are suitable for spam filtering applications, with Naive Bayes being more effective for this dataset, especially when fast computation and strong spam detection are priorities. Future work could explore additional machine learning models and larger datasets to further improve detection accuracy.

references

- Almeida, T.A., Hidalgo, J.M.G. and Yamakami, A., 2011. Contributions to the study of SMS spam filtering: new collection and results. Proceedings of the 11th ACM symposium on Document engineering, pp.259–262.
- Zhang, H., 2004. The optimality of naive Bayes. AAAI, 1(2), pp.3–10.
- Hastie, T., Tibshirani, R. and Friedman, J., 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd ed. New York: Springer.