Comparative Analysis Of Multinomial Naïve Bayes And SupportVector Classifier For SMS Spam Detection

Nicodemus Nonnious Mapogha
College of Aviation, Science And Technology
Lewis University
Romeoville Il, USA
nmapogha@lewisu.edu

***Abstract*—Despite the numerous advancements in today's digital age, SMS has remained an indispensable mode of communication. However, with its convenience comes the incessant problem of spam. This report presents the results of a programming project that aims to predict if an SMS is spam or not using two machine learning models: Multinomial Naive Bayes (MNB) and Support Vector Classifier (SVC). The project involves dataset preprocessing, model selection, and performance evaluation for these models. The results show that both models achieve high accuracy on the test set, but MNB has a slight edge over SVC in terms of precision and F1 score.**

Keywords—SMS Spam Detection, Machine Learning Models, Data Processing, Model Comparison, Performance Evaluation.

I. Introduction

Spam detection is an important and challenging task in natural language processing (NLP), as it can help protect users from unwanted or malicious messages. Spam messages can be classified into two types: email and SMS. Email spam is more common and well-studied, but SMS spam is also a serious problem that affects millions of mobile phone users around the world. One of the most popular and effective methods for spam detection is using machine learning models that can learn from labeled data and make predictions on new data. There are many machine learning models that can be applied to spam detection, such as Naive Bayes, Support Vector Machines, Decision Trees, Neural Networks, etc. However, different models may have different strengths and weaknesses depending on the characteristics of the data and the problem domain. Therefore, it is important to compare and evaluate different models to find the best one for a specific task. In this report, we focus on SMS spam detection using two machine learning models: Multinomial Naive Bayes (MNB) and Support Vector Classifier (SVC). MNB is a model based on probability that considers the words in a text message to be unrelated to each other and estimates the chance of a message being spam or not based on how often words appear. SVC is a model based on geometry that attempts to find a boundary (hyperplane) that divides the spam and non-spam messages in a high-dimensional space. Both models are widely used for text classification problems, such as sentiment analysis, topic modeling and document categorization. Through this project, we aim to compare and evaluate the performance of these two models on SMS sperm detection.

II. Dataset

I used a publicly available dataset from Kaggle, specifically the ”SMS Spam Collection” dataset. This dataset contains a collection of 5,571 text messages of which 4825 are labeled as ham (not spam) and only 747 are labeled spam. It includes 2 features representing the text and its label.

III. Data Preprocessing

The dataset required several preprocessing steps to ensure it was suitable for training machine learning models. Since the dataset did not contain any missing values, we went on to perform the following:

- Feature selection: The features were named 'V1' and 'V2' which had no meaning and so I renamed and rearranged them with 'text' representing the text messages and 'Clas' representing the labels of the text messages.
- Removing any encoding: I removed emoticons and any punctuation marks from the text messages as they may not carry much information for spam detection.
- Lowercasing: I converted all the text messages to lowercase as case sensitivity may not be relevant for spam detection.
- Tokenization: I tokenize each text message into a list of words using commas as a delimiter.
- Removing Stop Words and Lemmatization: I removed stop words and normalized words into their basic form by importing nltk library.
- Vectorization: I converted the text data to vectors by using CountVectorizer which counts how many times each word in the vocabulary appears in each text message.

After the above steps, I used the bag of words model to transform the 'text' column into a sparse matrix of token counts, which serves as the input features for the machine learning models. The output labels are the ham or spam classes for each text message

IV. Model Training

I split the preprocessed data into two subsets: 80% for training and 20% for testing. Using the sci-kit-learn Python library, I used the training set to train both models. The parameters for both models remained at default except for the regularization parameter C for SVC. I found that the optimal value of C for SVC is 0.85.

A. Model Evaluation

To assess the performance of the MNB and SVC models, I employed several evaluation metrics:

- Accuracy: The proportion of correctly classified text messages.
- Precision: The proportion of correctly classified spam text messages out of all text messages classified as spam.
- Recall: The proportion of correctly classified spam text messages out of all actual spam text messages.
- F1-Score: The harmonic mean of precision and recall.
- Confusion Matrix: A table that provides a more detailed breakdown of true positives, true negatives, false positives, and false negatives.

V. Results

A. Multinomial Naïve Bayes

The MNB model produced the following results:

TABLE I. The Confusion Matrix For MNB

	Predicted Ham (Not Spam)	***Predicted Spam***
Ham (Not Spam)	954	11
Spam	11	139

[a.] Table values are used to compute evaluation metrics.

- Accuracy: 0.98
- Precision: 0.98
- Recall: 0.98
- F1-Score: 0.98

B. Support Vector Clasiffication

The SVC model produced the following results:

TABLE II. The Confusion Matrix For SVC

	Predicted Ham (Not Spam)	***Predicted Spam***
Ham (Not Spam)	963	2
Spam	20	130

[b.] Table values are used to compute evaluation metrics.

- Accuracy: 0.98
- Precision: 0.98
- Recall: 0.98
- F1-Score: 0.98

VI. Discussion

The findings suggest that both the Multinomial Naive Bayes (MNB) and Support Vector Classifier (SVC) models are proficient in distinguishing spam SMS. Both models achieved commendable accuracy scores. However, relying solely on accuracy might not be the best approach for evaluating spam detection models, as it doesn’t account for the class imbalance issue or the implications of misclassification. Both models also exhibited similar weighted precision, recall, and F1-score values, which were notably high (98%). This indicates their effectiveness in accurately identifying spam text messages while minimizing both false positives and false negatives. Nevertheless, the SVC model outperformed the MNB model in terms of average precision, with only 2 messages incorrectly classified as spam compared to 11 by the MNB model. On the other hand, the MNB model had a superior average recall of 96%, with only 11 messages incorrectly classified as not spam compared to 20 by the SVC model. While both models performed admirably, the SVC model seems to be more suitable for this specific task of SMS spam classification due to its slightly higher precision in predicting spam. Although the primary objective of this project is to classify whether a text message is spam or not, it’s also crucial to minimize the number of messages incorrectly classified as spam to prevent the loss of important text messages.

VII. Conclusion

This project effectively showcased the use of machine learning techniques, specifically Multinomial Naive Bayes and Support Vector Classification, in the task of SMS spam identification. The data was meticulously cleaned, and features were extracted to make it suitable for the modeling process. The models were trained and assessed on the SMS Spam Collection Dataset, with accuracy, precision, recall, and F1-score serving as the evaluation metrics. Both models exhibited high accuracy in detecting SMS spam, however, the Support Vector Classification model slightly outperformed the Multinomial Naive Bayes model in terms of precision.

Acknowledgment

I would like to express my deepest gratitude to my instructors at Lewis University for their invaluable guidance and contributions to this project. I am particularly grateful to Dr. Mahmood Al-Khassaweneh, whose guidance in supervised machine learning has been instrumental in the development and success of my project. His insights and teachings have greatly enriched my understanding of different machine-learning algorithms.

References

[1] UCI Machine Learning, “SMS Spam Collection Dataset,” Kaggle. Accessed: Oct. 22, 2023. [Online]. Available: https://www.kaggle.com/datasets/uciml/sms-spam-collection-datase.

[2] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, 2009.

[3] S. Kaddoura, G. Chandrasekaran, D. E. Popescu, and J. H. Duraisamy, “A systematic literature review on spam content detection and classification,” PeerJ Computer Science, vol. 8, 2022. [Online]. Available: https://doi.org/10.7717/peerj-cs.830.