A Hybrid Approach to Email Spam Detection: Random Forest and Sentiment Analysis

ShuyanLiu
Division of Programs in Business
School of Professional Studies
New York University
New York, New York
Email:sl10158@nyu.edu

EleftheriaKPissadaki
Division of Programs in Business
School of Professional Studies
New York University
New York, New York
Email:ep3041@nyu.edu

ThomasMSchmidt
Division of Programs in Business
School of Professional Studies
New York University
New York, New York
Email:tms493@nyu.edu

***Abstract*—Email spam detection is a persistent problem in onlinecommunication,andtraditionaldetectionmethodsstruggle tokeepupwithevolvingspamtactics.Thispaperproposesahybridapproachtothedetectionofemailspam,combining the strengths of random forest classification and sentiment analysis. Our approach leverages the robust feature selectionand classification capabilities of random forest to identify spam patterns, while incorporating sentiment analysis to capture the nuances of language used in spam emails. Our hybrid approach achieved an accuracy of 94.54%, outperforming the benchmark model Na¨ıve Bayes. Our results show that integrating sentiment analysis with random forest classification can effectively combat email spam, making it a powerful tool for spam detection in the modern email communication environment.**

I. Introduction

In the realm of digital communication, email remains a crucial medium, facilitating personal, academic, and professional exchanges throughout the world. However, this ubiquity also makes email a prime target for spam, which not only disrupts communication,butalsoposessecurityrisks.Traditionalspam detection methods, such as rule-based filtering and Na¨ıve Bayes classifiers, have been widely employed, but often fall short in dealing with the sophisticated and ever-evolving strategies employed by spammers. Consequently, there is a pressing need for more advanced and adaptive spam detection techniques.

While the Na¨ıve Bayes algorithm has shown efficacy in spam detection, achieving accuracy rates such as 91.13% and 82.54% in studies by Rusland et al. [1] and Zhang et al. [2],its performance can be limited by its simplicity. Na¨ıve Bayes assumes feature independence and often struggles with spam emails that cleverly blend legitimate content with deceptive cues. This simplicity, while beneficial for straightforward applications, can hinder its ability to adapt to sophisticated spammingtechniquesthatexploitcorrelationsbetweenfeatures.

This paper introduces a hybrid approach to enhance the accuracyandefficiencyofemailspamdetection.Byintegrating the robust classification capabilities of Random Forest withthe nuanced language understanding afforded by sentiment analysis, this method aims to identify and filter spam emails moreeffectively.UnlikeNa¨ıveBayes,RandomForestdoes not assume feature independence and can handle the interplay betweenvariablesmoreeffectively.Itconstructsmultipledecisiontreesonvarioussubsetsofthedatasetandaggregatestheir predictions, leading to more robust and less biased results. Sentiment analysis, on the other hand, provides insights into the emotional and psychological constructs within the text, which are often manipulated by spam content to mislead or entice the recipient.

Our hybrid model not only addresses the limitations of traditional algorithms, but also sets a new standard in spam detection by achieving an accuracy of 94.54%, thus significantlyoutperformingthebenchmarkmodelNa¨ıveBayes.This paperdetailsthedevelopment,implementation,andevaluation of our approach, highlighting its potential as a powerful tool against the persistent challenge of email spam in modern communication environments.

II. ProposedMachineLearningModels

A. Na¨ıveBayesClassifier

TheNa¨ıveBayesclassifierisaprobabilisticmachinelearningmodelthatiswidelyusedforemailspamdetectiondue to its simplicity and effectiveness. It is based on Bayes' Theorem, which uses the probabilities of events to make predictions. This model assumes that the presence (or absence)ofaparticularfeatureinaclassisnotrelatedto the presence (or absence) of any other feature, known as conditional independence. Despite this simplification, Na¨ıve Bayes can perform remarkably well and is especially fast for the training and prediction phases. It is known to outperform evenhighlysophisticatedclassificationmethods.T.M.Ma,K. YAMAMORI and A. Thida [3] The mathematical expression can be formulated as following:

$$P(c|x)= \frac{P(x|c)P(c)}{P(x)} \qquad (1)$$

where,

- $P(c \mid x)$ is the posterior probability of class c given predictor(s) x. After observing the email's content x, this posteriorprobabilityhelpstodeterminewhethertheemail is more or less likely to be classified as spam.

- $P(x \mid c)$ is the likelihood, which is the probability of observingthepredictor(s) $x$giventhatthe classis$c$.This is the probability of observing the specific characteristics of an email x assuming that it is spam c.
- $P(c)$ is the prior probability of class c, indicating how frequent the class c is in the data set before observing x. In our task this represent the overall frequency of spam emails in the dataset.
- $P(x)$is the marginal likelihood or evidence, which is the probability of observing predictor(s) x across all classes. This represents the probability of observing the specific features x in any email, regardless of whether it is spam or not.

WeemploytheNa¨ıveBayesmodelagainstwhichwecom- pare the performance of our proposed hybrid spam detection apprach.

B. RandomForestClassifier

Random Forest (RF) is an ensemble learning method, renownedforitshighaccuracyandrobustness.Developedby L. Breiman [4], this algorithm improves decision making by combiningthepredictionsofmultipledecisiontreesintoafinal output.Eachtreeintheensembleisconstructedfromadistinct bootstrap sample of the data. During the construction of these trees, the nodes are split using the optimal split selected froma randomly chosen subset of the features. This method not only leverages the strength of multiple learning models, but also introduces randomness into the model selection process, significantly reducing the risk of overfitting. Its speed and efficiency when applied on large datasets, it doesn't overfit,no presumptions on the distribution of the data are needed. L. Guo et al. [5]

III. METHODOLOGY

*1) Dataset:*Theempiricalevaluationofourspamdetection model uses three distinct data sets to ensure a comprehensive assessment across various types of emails. The TREC 2006 SpamDatacomprise53,668emails,splitbetween29,923ham and 23,745 spam messages, providing a substantial volumefor both training and testing our algorithms. Following this,the TREC 2007 Spam Data includes 17,309 emails, with a division of 12,508 ham and 4,801 spam messages. Lastly, we includeaSpamEmailDatasetfromKaggle,whichconsists of 5,127 emails, closely balanced with 2,259 ham and 2,868 spammessages.Thisbalanceddatasetiscriticalfortestingthe model's effectiveness in scenarios where spam and ham are nearly equally represented. The varied nature and complexity of these datasets are instrumental in evaluating the generalizabilityandreliabilityoftheproposedspamdetectionsystem.

2) Data Preprocessing: Effective data preprocessing is crucial for minimizing noise and enhancing the performanceof the model. Our preprocessing pipeline incorporates several steps to standardize and refine the input data, ensuring it is optimallypreparedforthesubsequentmodelingphases.These steps include:

Lowercasing: All text data is converted to lowercase to maintain consistency across the dataset, eliminating variations caused by case differences.

Removal of Redundant Prefixes: We remove specific prefixes, such as 'subject:', which could introduce bias if left within the text, as these elements do not contribute to distinguishing between spam and legitimate emails.

Numeric Replacement: Numeric values within the texts are replaced with a placeholder token. This approach prevents numbers from skewing the model's learning process, as their presence varies widely across emails.

Removal of Stopwords: Stopwords, which are commonly occurring words that offer little analytical value, are removed to sharpen the focus on more meaningful terms. This step is critical in spam detection as it reduces the data dimensionality andemphasizes keywordsthat aremore likelyto beindicative of spam or non-spam emails.

Tokenization: Text data is broken down into individual words or tokens.

Lemmatization: Words are reduced to their base or dictionary forms using lemmatization. This process helps in consolidating different forms of the same word, ensuring that variations in tense or plurality do not affect the analysis.

Each of these preprocessing steps is designed to refine the dataset,reducingredundancyandemphasizingfeaturesthatare most informative for spam detection. This careful preparation is essential for building a robust model capable of effectively identifying spam emails.

3) Feature Extraction: TF-IDF Transformation: We employed Term Frequency-Inverse Document Frequency (TF-IDF) to transform the text data into a numerical format that is more amenable to machine learning algorithms. As noted by J.Ramos[6]inhis2003study,TF-IDFweighsthewords'frequenciesacrossdocumentsagainsttheirdistributionacrossthe entirecorpus,highlightingwordsthatareparticularlypertinent toindividualdocumentsmorethanthosecommonthroughout. This method enhances the model's ability to discern and learn from the patterns in the text data.

4) Sentiment Analysis: Sentiment analysis plays a pivotal roleinourspamdetectionsystem,significantlyenhancing its capability to differentiate between legitimate emails and spam. This technique leverages natural language processing (NLP) to assess the emotional content of email messages. Spam emails often employ distinct emotional cues, such as urgency or exaggerated positivity, to manipulate the recipient. These cues serve as significant indicators for distinguishing spam from legitimate emails. To quantify these emotional tones, we utilize the TextBlob library, which assigns polarity scorestothetext.Thesescoresrangefrom-1.0,indicating a very negative sentiment, to 1.0, reflecting a very positive sentiment. For integration with our Na¨ıve Bayes model, we adjustthese scores from the original range of $[-1,1]$ to [0,2]. This adjustment better aligns with the model's requirements for handling input values, ensuring that the sentiment scores are effectively utilized to enhance spam detection accuracy.

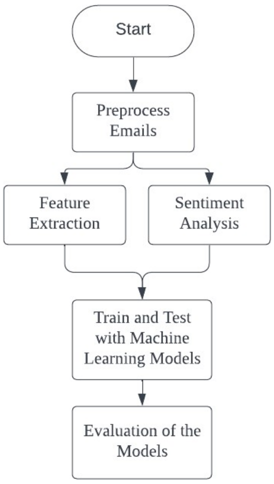

Fig.1.ModelBuildFlowChart

5) Experimental Setup and Model Comparison: To rigorously evaluate the efficacy of our spam detection system, we will conduct a series of experiments comparing four different models the model building process is shown in Fig. 1.

- Na¨ıve Bayes (NB): This baseline model uses standard spam detection techniques without sentiment analysis, relying primarily on TF-IDF scores.
- Random Forest (RF): As another baseline, this model applies a Random Forest algorithm, using the same TF-IDF features.
- Na¨ıve Bayes with Sentiment Analysis (NB-SA): This model enhances the Na¨ıve Bayes approach by incorporating sentiment analysis, adding sentiment scores as additional features to the model to see if recognizing emotional cues improves spam detection.
- Random Forest with Sentiment Analysis (RF-SA): This model combines Random Forest with sentiment analysis. We hypothesize that this model will perform the best due toitsabilitytoutilizeboththestructureddecision-making of Random Forest and the nuanced understanding of text sentiment.

Each model will be assessed using standard performance metrics in machine learning. A confusion matrix will serve as the basis for evaluating each proposed approach in terms of false positive, false negative, accuracy, precision, and recall.

IV. RESULTSANDDISCUSSION

TABLE 1 presents detailed results of our proposed models. demonstrating that the baseline Random Forest model (RF) significantly outperforms the Naive Bayes model (NB). The NB model displays a notably high False Positive Rate (FPR) of 38.91%, in contrast to the RF's 10.67%. In applications such as email spam detection, such a high FPR can lead to considerable user inconvenience by misclassifying legitimate emailsasspam.Moreover,theRFmodelexhibitssuperior accuracy, achieving 87.32% compared to 66.67% for the NB model.

TABLEI
CLASSFICATIONRESULTFORPROPOSEDMODELS

Model	FP-Rate	FN-Rate	Accuracy	Precision	Recall
NB	26.25%	38.91%	66.67%	68.10%	66.67%
RF	15.23%	10.67%	87.32%	87.31%	87.32%
NB-SA	23.55%	21.24%	77.80%	78.10%	77.80%
RF-SA	7.59%	3.69%	94.54%	94.56%	94.54%

Integration of sentiment analysis markedly improves both models, with the Random Forest with Sentiment Analysis model (RF-SA) showing exceptional performance improvements, reducing the False Negative Rate to 3.69% and boosting accuracy to 94.54%. These enhancements suggest that while sentiment analysis can elevate the performance of basic models,theselectionoftheunderlyingalgorithmispivotal in maximizing the utility of advanced feature processing techniques.

TheimprovementfromNBtoNB-SAandRFtoRF-SA highlights the significant role of sentiment analysis in capturing nuanced data details, thereby enhancing predictive accuracy.

The Receiver Operating Characteristic (ROC) curves, as illustrated in Fig. 2, provide additional insights into the models'abilitytodiscriminatebetweenclasses,affirming the superior discriminative power of the RF-SA model. The blue curve represents the ROC curve for the Naive Bayes-Sentiment Analysis (NB-SA) classifier, with an area underthe curve (AUC) of 0.81. This value indicates the classifier's abilitytodistinguishbetweentheclasses,where1.0represents perfect classification and 0.5 represents a random guess. The green curve represents the ROC curve for the Random Forest-Sentiment Analysis (RF-SA) classifier, with an AUC of 0.98, which suggests a very good model performance, much better thantheNB-SAclassifier.Thedasheddiagonallinerepresents a random classifier. A good classifier stays as far away from this line as possible (towards the top left corner).

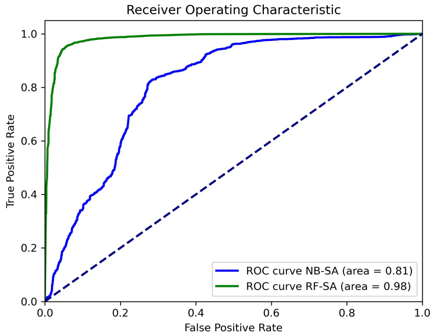

Fig.2.ROCCurvesforNB-SAandRF-SAModels

V. CONCLUSION

Thisstudyvalidatestheeffectivenessofahybridemailspam detectionapproachthatintegratesRandomForestclassification with sentiment analysis. This combination leverages the powerful feature selection capabilities of Random Forest and the nuanced detection offered by sentiment analysis to effectively combat the sophisticated tactics of modern spam emails. Our findings indicate that this method excels at identifying subtle linguisticindicatorstypicalofadvancedspamcampaigns,thus minimizing false positives and enhancing the reliability of the detection system.

Future research should focus on refining the integration of sentiment analysis within the classification process, exploring the potential of more granular linguistic features such as syntaxandcontextualsemanticstofurtherenhancethemodel's accuracy.Additionally,experimentingwithdeeplearningtechniques, particularly those specializing in natural language processing, could offer new insights and improvements in spam detection capabilities.

Subsequent studies should also assess the scalability and efficiency of this hybrid model across larger and more diverse datasets, as well as under varying operational conditions, to confirm its effectiveness in a real-world environment. Such evaluations would provide deeper insights into the model's practical applications and limitations, facilitating its broader adoption and implementation in spam detection systems.

REFERENCES

[1] N. F. Rusland, N. Wahid, S. Kasim, and H. Hafit, ”Analysis of Na¨ıveBayes algorithm for email spam filtering across multiple datasets,” in*IOP Conference Series: Materials Science and Engineering*, vol. 226,no. 1, Aug. 2017, Art. no. 012091.

[2] H.Zhang,N.Cheng,Y.Zhang,andZ.Li,”LabelflippingattacksagainstNa¨ıve Bayes on spam filtering systems,” *Applied Intelligence*, vol. 51,no. 12, pp. 4503–4514, 2021.

[3] T.M.Ma,K.Yamamori,andA.Thida,”AComparativeApproachto Na¨ıve Bayes Classifier and Support Vector Machine for EmailSpam Classification,” in *2020 IEEE 9th Global Conference on Con-sumer Electronics (GCCE)*, Kobe, Japan, 2020, pp. 324-326, doi:10.1109/GCCE50665.2020.9291921.

[4] L. Breiman, ”Random Forests,” *Machine Learning*, vol. 45, pp. 5–32,2001.

[5] L. Guo, N. Chehata, C. Mallet, and S. Boukir, ”Relevance of airbornelidar and multispectral image data for urban scene classification usingRandom Forests,” *ISPRS Journal of Photogrammetry and Remote Sens-ing*, vol. 66, no. 1, pp. 56–66, 2011, Elsevier.

[6] J. Ramos, ”Usingtf-idf to Determine Word Relevance in DocumentQueries,” in *Proceedings of the First Instructional Conference onMachine Learning*, vol. 242, no. 1, pp. 29-48, Dec. 2003.