2384: PHISHING EMAIL DETECTION USING MACHINE LEARNING

YOGA SHRI MURTI, 1191100796
SUPERVISOR: DR.NAVEEN PALANICHAMY MODERATOR: DR.TAN SAW CHIN
MMU MULTIMEDIA UNIVERSITY | FACULTY OF COMPUTING & INFORMATICS

Abstract

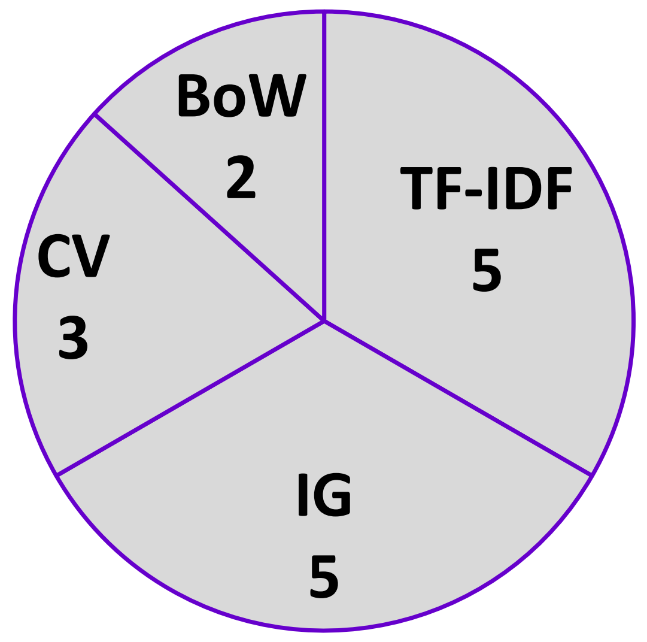
- Phishing emails pose significant threats, and their detection is crucial for digital communication security.
- Current machine learning methods for detecting phishing emails are either slow or ineffective.
- This study develops a reliable phishing email detector using a hybrid machine classifier with TF-IDF feature extraction.
- The proposed tuned hybrid model achieves high performance, with 93.8% accuracy, 1.0 precision, 87.5% recall, and 94% F1-score.
- The study highlights the value of machine learning for detecting phishing emails and emphasizes the benefits of using a combination of models for improved performance.

Problem Statement & Objective

How can we enhance phishing email detection using a hybrid machine learning model?

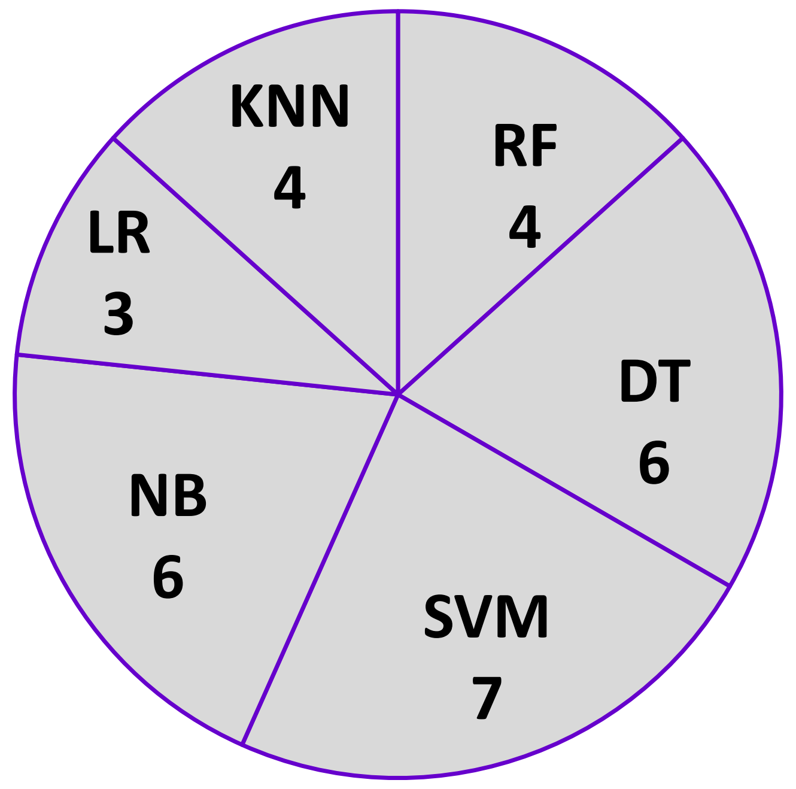
- By combining various machine learning techniques and models, the hybrid approach effectively detects phishing emails.
- This integration uplift the capability in categorizing the emails correctly, leading to enhanced cybersecurity measures and time saving.

Literature Review



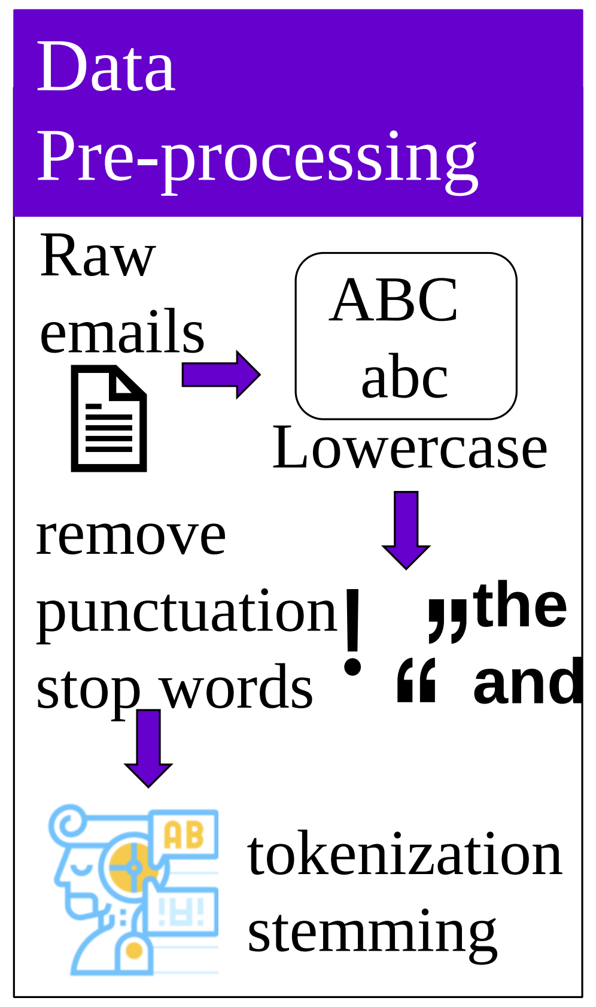
- Summary of feature extraction used in reviewed papers.
- As a result, this project will use TF-IDF and count vectorization.

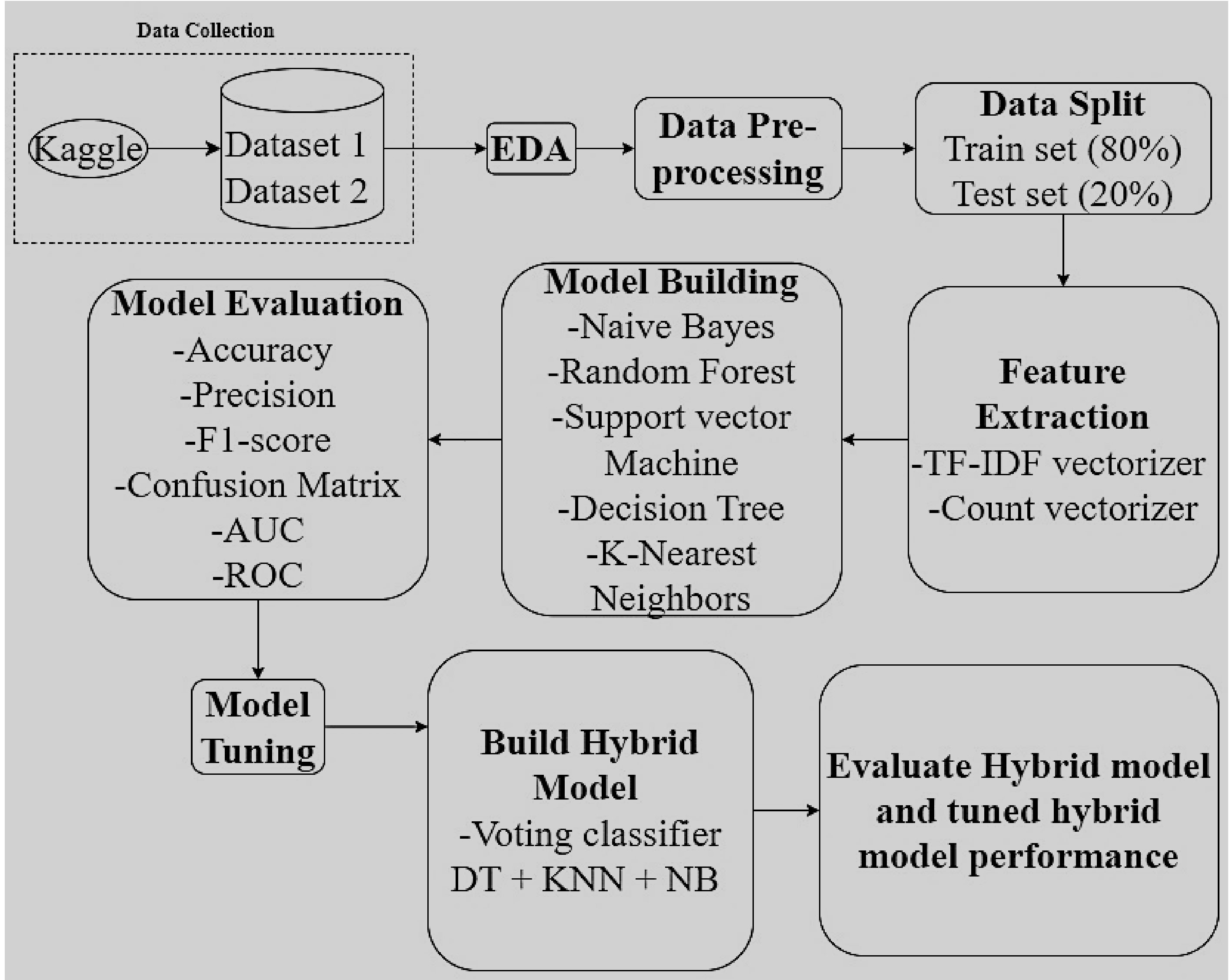
- Summary of classifier used in reviewed papers.
- NB, RF, and DT are recommended methods in prior studies, as they perform well and are efficient in terms of time.



Research Methodology

Figure illustrates the methodology flow for proposed approach of this project.





Implementation & Evaluation

- Experiment A

Model	Time taken to train	Time taken to test	Train accuracy	Test accuracy
NB	0.31s	0.06s	96.754	96.669
KNN	0.11s	152.58s	98.576	98.316
RF	133.61s	1.99s	99.999	99.451
SVM	423.44s	313.3s	98.187	98.130
DT	9.4s	0.06s	100.0	99.104

- Experiment B

Model	Accuracy TF-IDF	Accuracy CV	Precision TF-IDF	Precision CV
NB	0.9375	0.8750	1.0000	1.0000
KNN	0.8125	0.5000	1.0000	0.5000
RF	0.7500	0.6875	0.6667	0.6353
SVM	0.7500	0.6250	1.000	0.5833
DT	0.8125	0.8125	0.7778	0.7778

- Experiment C

Model	Accuracy	Precision	Recall	F1-score
Hybrid Model with (TF-IDF) DT + KNN + NB	0.8125	1.0	0.625	0.7692
Tuned Hybrid Model	0.9375	1.0	0.875	0.9333

Conclusion & Future Work

- Proposed tuned hybrid model could be a promising tool in the detection of phishing emails.
- It combines individual model strengths making predictions, reducing overfitting, and improving generalization.
- Future work in this field could focus on continuous research, and creating it into a public detection app.

References

- Adi Wijaya, & Achmad Bisri . (2016). Hybrid Decision Tree and Logistic Regression Classifier for Email Spam Detection. 2016 8th International Conference on Information Technology and Electrical Engineering (ICITEE), Yogyakarta, Indonesia (p. 4). IEEE Xplore.
- J. Vijaya Chandra, Narasimham Challa, & Sai Kiran Pasupuletti. (2019, october). Machine Learning Framework To Analyze Against Spear Phishing. 8(12). doi:10.35940/ijitee.L3802.1081219
- Lew May Form, Kang Leng Chiew, San Nah Sze, & Wei King Tiong. (2022, 9 25). Phishing Email Detection Technique by using Hybrid Features. 5.
- Jawale, D. S., Diksha S. Jawale , Kalyani R. Shinkar , & Kalyani R. Shinkar . (2018). Hybrid spam detection using machine learning. International Journal of Advance Research, Ideas and Innovations in Technology, 4(2), 1-6.

Publications

- Machine learning algorithm for phishing e-mail detection published in Journal of system and management science.
- Improving Phishing Email Detection Using Hybrid Machine Learning Approach paper was approved in Conference on Computer , Information Technology and Intelligent Computing 2023 (CITIC 2023).

Acknowledgement

I am immensely grateful for the opportunity to undertake my Final Year Project ,thanks to God. Special appreciation goes to my supervisor, Dr. Palanichamy Naveen. Then , I'm grateful for my family and friends support. Lastly, thank you dear self for never quitting.

SCAN ME