



Task 8 – Critique of the Statistical Analysis

Prepared by

ALI MMADI, PhD student-Midocean University



Literary Criticism (2)

Task 8

INTRODUCTION

Chosen topic for Task 8:

“Statistical Evaluation of Phishing Email Detection Models Using Natural Language Processing (NLP)”

(Fictitious but realistic article, based on common cybersecurity work, 2023)

INTRODUCTION

Hello everyone,
I am Ali Mmadi, a PhD student in Information Technology, specializing in Cybersecurity.

In this eighth assignment of the Critical Studies 2 module, I will analyze and critique the statistical method used in a research study titled:

“Statistical Evaluation of Phishing Email Detection Models Using NLP Techniques” (2023)

This study aims to compare different machine learning models applied to phishing emails, using statistical methods to evaluate their performance.

Statistical Evaluation of Phishing Email Detection Models Using NLP Techniques

1) Relevance of the choice of statistical methods

In the study, the authors used several statistical analyses:

Descriptive statistics of the data (mean, variance, token distribution).

Confusion matrices (True Positive, False Positive, etc.).

Comparative statistical measures: accuracy, recall, F1-score.

ANOVA hypothesis test to compare the performance of different NLP models.

Critical analysis:

- ✓ The statistical methods are generally appropriate for the type of data (quantitative).
- ✓ The use of ANOVA is relevant for comparing multiple learning models on the same dataset.
- ✓ Descriptive analysis is well used to explore the characteristics of the dataset (email length, frequency of suspicious words...).

✗ Weak point:

The researchers did not explain why ANOVA was chosen instead of a non-parametric test like Kruskal-Wallis, even though the normality of the data was not demonstrated.

This raises doubts about the statistical validity of the results.

Normality Test

Classifier	statistic	<i>p</i> -value
BERT	0.946	0.107
SVM	0.946	0.106
RF	0.942	0.062
Logistic Regression	0.951	0.254

BH Score normality using Shapiro-Wilk

* Significant at $p < .05$

2) Type of data and suitability for analysis

The dataset is entirely quantitative, consisting of:

- 10,000 emails (6,000 legitimate, 4,000 phishing)
- Numerical variables extracted by NLP:
 - ❑ tf-idf,
 - ❑ sentence length,
 - ❑ similarity score,
 - ❑ number of URLs,
 - ❑ presence of suspicious words.

Critical analysis:

- ✓ The quantitative data lend themselves well to the statistical analysis used.
- ✓ The authors clearly explained the nature of the data.

✗ Limitation:

The study does not mention any treatment for class imbalance, which is essential in cybersecurity (there are more legitimate emails than malicious ones).
The absence of this step can skew the statistics.

3) Use of descriptive and inferential statistics

The authors used:

Descriptive statistics:

Mean

Median

Standard deviation

Frequency distribution

Relevant because:

✓ This allows understanding the structure of the emails, especially for NLP features.

✓ The authors provide good statistical visualization: histograms, box plots.

Inferential statistics:

ANOVA to compare the performance of BERT, SVM, Random Forest, and Logistic Regression.

95% confidence interval.

Critical analysis:

- ✓ The use of ANOVA makes sense for comparing multiple models.
- ✓ Confidence intervals reinforce statistical significance.

✗ Weaknesses:

- No normality test (Kolmogorov-Smirnov or Shapiro-Wilk).
- No post-hoc test (Tukey HSD) to explain which pairs of models are different.
- No justification for the $\alpha = 0.05$ threshold.
 - The statistics are correctly applied, but incomplete at a doctoral level.

4. Qualitative data processing

The research contains no qualitative data, as it is based solely on numerical variables generated by NLP.

Critical analysis:

- ✓ It is consistent: automatic phishing email detection is a quantitative task.
- ✗ However, a brief qualitative analysis of the fraudulent emails (intentions, motives, linguistic patterns) would have enriched the understanding of the phenomenon.



Critical analysis:

- ✓ The software is suitable for quantitative analysis in cybersecurity.
- ✓ The statistical metrics are relevant.
- ✗ However, the authors do not explain how the hyperparameters were chosen,
- ✗ nor whether they performed cross-validation.

Without these details, the credibility of the results is less solid.

General limits of statistical analysis

Here are the main identified limitations:

- No sensitivity analysis on hyperparameters.

No robustness testing has been carried out.

⚠ The potential non-normality of the data calls into question the use of ANOVA.

⚠ No correlation test between NLP variables (e.g., Pearson or Spearman).

CONCLUSION

In conclusion, the statistical analysis used in this study is overall relevant, well-structured, and consistent with the objectives. The descriptive and inferential methods are suitable for the type of quantitative data.

However, several methodological aspects need improvement, notably:

- the justification of statistical tests,
- the verification of normality conditions,
- the addition of post-hoc tests,
- and better management of unbalanced data.

Thus, even though the study is solid, it remains statistically improvable to reach a doctoral level.

Thank you for your attention.

1. **hapiro, S. S., & Wilk, M. B. (1965).**
An analysis of variance test for normality (complete samples). Biometrika, 52(3–4), 591–611.
→ Basis of the Shapiro-Wilk normality test.
2. **Kruskal, W. H., & Wallis, W. A. (1952).**
Use of ranks in one-criterion variance analysis. Journal of the American Statistical Association, 47(260), 583–621.
→ Foundational reference for the Kruskal–Wallis non-parametric test.
3. **Fisher, R. A. (1925).**
Statistical Methods for Research Workers.
→ Classical reference for ANOVA methodology.
4. **Pedregosa, F. et al. (2011).**
Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825–2830.
→ Main framework used in research evaluating ML models, including NLP phishing detection.
5. **Chiew, K. L., Yong, K. S. C., Tan, C. L., & Wong, K. (2019).**
A review of phishing attacks and detection mechanisms. Computers & Security, 83, 11–28.
→ Major reference for phishing detection research.
6. **Sahingoz, O. K., Buber, E., Demir, O., & Diri, B. (2019).**
Machine learning based phishing detection from URLs. Expert Systems with Applications, 117, 345–357.
→ Frequently cited study comparing ML models for cyber-attack detection.
7. **Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019).**
BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.
→ Reference for NLP model used in phishing text detection.
8. **Abu-Nimeh, S., Nappa, D., Wang, X., & Nair, S. (2007).**
A comparison of machine learning
A comparison of machine learning techniques for phishing detection.
→ Historical but foundational benchmarking study.
9. **Goodman, S. N. (1999).**
Toward evidence-based statistics: the P-value fallacy. Annals of Internal Medicine, 130(12), 995–1004.
→ Useful reference for criticism of p-values, often cited in ANOVA critiques.
10. **Ben-David, A., Sterling, L., & Shapira, B. (2020).**
Statistical pitfalls in machine learning research and evaluation. Journal of Artificial Intelligence Research.



مجموعة جامعة
الجميع الذكية
للتعليم القابضة

MIDOCEAN
UNIVERSITY



Thank you!!!

