Science World Journal Vol. 20(No 2) 2025 https://dx.doi.org/10.4314/swj.v20i2.20
www.scienceworldjournal.org
ISSN: 1597-6343 (Online), ISSN: 2756-391X (Print)
Published by Faculty of Science, Kaduna State University

Full Length Research Article

COMPARATIVE ANALYSIS OF FEATURE EXTRACTION TECHNIQUES FOR SPAM DETECTION

*[1]Gbenga O. Ogunsanwo, [1]Blessing C. Ngoka, [1]Olumiywa O. Alaba, [2]Ayokunle A. Omotunde

[1]Department of Computer Science, College of Science and Information Technology, Tai Solarin University of Education, Ogun State, Nigeria
[2]Department of Computer Science, Babcock University, Ilisan Remo , Ogun State, Nigeria

*Corresponding Author Email Address: ogunsanwogo@tasued.edu.ng

ABSTRACT
The advent of smartphones has tremendously increased the spam rate in the communication sector. Developing a predictive model for spam detection plays a crucial role in enhance online security, improving user experience and protecting businesses from various risk that comes with spam. Feature extraction (FE) is a very important stage in increasing the accuracy of the model. This study, therefore, developed a comparative study of five FE techniques on a spam dataset. The study used dataset from Kaggle repository which contain 5,574 SMS messages in English tagged as ham (legitimate) or spam. The study employed five FE techniques which are: BoW, PCA, TF-IDF, N Gram and BERT with two classifier which are SVM and LR. The results pointed out that BERTS FE usually lead to the highest accuracy for the experiment carried out, while both SVM and LR achieved their best accuracy of 0.989 and 0.990 respectively. The study concluded that the accuracy results highlight the importance of choosing appropriate feature extraction techniques. The study recommends that careful selection of FE methods will optimize the model performance. Further works can be done with different dataset with different FE techniques and different Deep learning algorithm.

Keywords: feature extraction, spam, detection, classifier

INTRODUCTION

Machine learning techniques are gaining more attention in decision-making in various sectors due to its ability to draw out valuable information from large dataset. When developing predictive models, feature extraction (FE) is very important as it helps to increase the accuracy of the predictive model. FE also plays a crucial role in developing a natural language processing (NLP) tasks in the way it assist in transforming raw text data into numerical representation that machine learning (ML) algorithms can easily use or process (Manning et al., 2008). There are numerous FE techniques, the choice of FE can importantly affect the performance of ML tasks such as classification (Sebastiani, 2002) churn prediction (Ogunsanwo, 2025) sentiment analysis (Pang & Lee, 2008), Movie prediction (Omotunde et al.,) and Air quality index (Ogunsanwo et al., 2025).

Numerous feature extraction methods have been developed, each with its strengths and weaknesses. Traditional approaches like Bag-of-Words (BoW) (Salton & McGill, 1986) and Term Frequency-Inverse Document Frequency (TF-IDF) (Sparck, 1972) focused on word frequencies and document-level statistics. Many authors have worked in FE analysis such as: Fadhel et al (2025) carried out a comparative study of FE methods for multiple option classification using dataset from Twitter on airline and Borderland game review. The results shows that TF-IDF approach explored confirmed that ETC achieved the highest accuracy rate of 94% and 90% for Airline and Borderland. The study concluded that Fastext with the X-Gboost classifier outperform all other techniques achieving 94% accuracy on the Airline.

Hemdanou et al., (2024) carried out a comparative analysis of feature selection and extraction methods for student performance prediction across different machine learning models. The study employed two feature extraction techniques which are PCA and Variation Autoenconder. The study also employed Machine learning and Deep learning classifiers which are Decision Tree, Random forest , KNN , SVM ,DNN and RNN-GRU. The study revealed that the Transformer Model performed better than the other models developed in terms of the validation metrics used which are MAE and RMSE.

Mohtasham et al. (2024) carried out a study on comparative analysis of FE techniques on COVID-19 dataset. The study employed 13 ML models to check the effectiveness of these FE techniques based on classification accuracy. The results shows that the Hybrid Boruta-VI model combined with the RF algorithm demonstrated the best performance in achieving accuracy of 0.89, an F1 score of 0.76 and AUC value of 0.95.

Shuai1 et al.,(2020) carried out a study on feature extraction methods namely: BoW , W2V and BERT for automated ICD coding. The study revealed that BERT variants with the whole network was optimal for tasks involving only frequent code, majorly code covering unspecified disease while BoW turned out to be best for tasks involving both frequent and infrequent codes .

Wamidh et al., (2022) carried out a study on Feature Extraction Methods: The study divided the features into four groups; Geometric features, Statistical features, Texture features and Color features. The study makes a comparison among them by using two types of images (Face image and Plant image). Plant image has the best accuracy at 98%.

Calesella et al. (2021) carried out a study on comparative analysis of FE for prediction of neuropsychological scores from functional connectivity data of stroke patients. The study accessed four well known FE techniques which are: Principal Component Analysis (PCA), Independent Component Analysis Dictionary Learning (DL) and NonNegative Matrix Factorization (NNMF). The study concluded that PCA and ICA were best at extracting representative features and this results shows that features extracted by PCA and ICA were found to be the best predictors of neuropsychological scores for all the considered domain. However, they often overlook the semantic relationships between words, which are captured by more recent techniques like word embedding like Word2Vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014). Additionally, advanced methods like BERT (Devlin et al., 2019) have emerged, leveraging deep learning architectures to generate contextualized word representations.

Science World Journal Vol. 20(No 2) 2025 https://dx.doi.org/10.4314/swj.v20i2.20
www.scienceworldjournal.org
ISSN: 1597-6343 (Online), ISSN: 2756-391X (Print)
Published by Faculty of Science, Kaduna State University

Given the diversity of available techniques, it is essential to understand their relative performance across different datasets and classification algorithms. This study aims to provide a comparative analysis of the five prominent feature extraction techniques: BoW, TF-IDF, Word2Vec, GloVe, and BERT. We will evaluate their performance on distinct datasets, chosen to represent different text characteristics and domains. To ensure a comprehensive assessment, we will employ two widely used classifiers: Support Vector Machines (SVM) (Cortes &Vapnik, 1995) and Random Forest (Breiman, 2001), known for their effectiveness in various NLP tasks (Fernández-Delgado et al., 2014).

By conducting this comparative study, we seek to address the following research questions: .How does the five feature extraction techniques compare in terms of classification accuracy on the two classifiers? Are there significant differences in performance between the two classifiers when using the same feature extraction method?

.Are there any computational trade-offs associated with the different feature extraction methods? The main contributions of this paper to the advancement of knowledge includes: Comparative analysis of five feature extraction techniques, implement the five feature extraction techniques, hybridizing the feature extraction with PCA, implement two classifiers on the extracted features, numerous evaluation of the different model parameters and classification accuracy

MATERIALS AND METHODS

The data acquisition for this study was downloaded from Kaggle repository. It contains one set of SMS messages in English of 5,574 messages, tagged according being ham (legitimate) or spam. The study employed five FE techniques which are : BoW, PCA, TF-IDF, N Gram and BERT with two classifier which SVM and LR. The Figure 1 depicted the flow of the work done

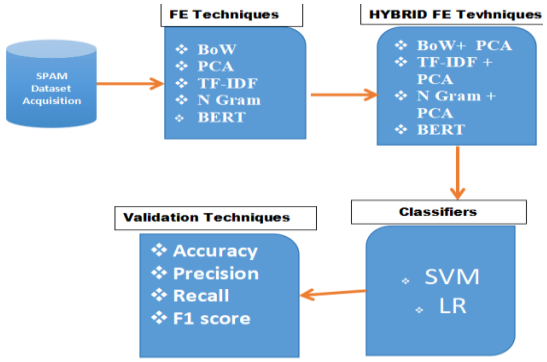

Figure 1 Flow Work

BoW

The BoW is commonly used feature extraction techniques used in Natural language processing (NLP) for turning data as numerical features. BoW treats a document as an unordered collection of words, neglecting grammar and word order. It is very easy to implement and understand. It major on the frequency of words inside a documents this might resulted in the loss of semantic information (Zhang et al., 2010)

PCA

PCA is a unique technique for dimensionality reduction and feature extraction, offering valuable insights in data analysis and machine Learning (ML). It aims at identifying the main component which are new variables that capture the highest variance in the data. It help to reduce the complexity while protecting relevant information (Jolliffe & Cadima 2016).

TF-IDF

TF-IDF is one the most popularly used methods for text analysis, even with the advancement in Deep learning. It measure the importance of word inside a document with reference to a group of documents (corpus). The main principle is weighing terms based on their frequency in a document and their inverse frequency across the group of document. This is done by using two component which are TF and IDF. The TF measures how regular a term occurs in a specific document while the IDF measures the importance of a term across the entire corpus (Manning et al., 2008).

N gram

The N-gram FE is one of the technique used in NLP to represent text data by examining the sequence of N consecutive tokens (words). It builds upon the BoW model by inserting a local word order information which can use more context and meaning when compared to individual words (Cavnar & Trenkle 1994).

BERT

BERT (Bidirectional Encoder Representation from Transformer) is a good language model that can be employed for different NLP tasks. It has the ability to extract feature from text that can be used to train machine learning model (Devlin et al., 2018) The mathematical representation of BERT is seen in Equation (1).

$$\text{BERT}(y) = \text{TransformerEncoder}(\text{Embedding}(x)) \quad (1)$$

where:
y is the input text.
Embedding is the function that converts tokens to embeddings.
Transformer Encoder is the stack of Transformer encoder layers.

SVM
LR

$$\text{Accuracy} = \text{TN} + \text{TP} / (\text{TN} + \text{TP} + \text{FP} + \text{FN}) \quad (2)$$

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP}) \quad (3)$$

$$\text{Sensitivity (recall)} = \text{TP}/ \text{TP} + \text{FP} \quad (4)$$

$$\text{Specificity} = \text{TN} / (\text{TN} + \text{FP}) \quad (5)$$

$$\text{F1 Score} = 2 * \text{Precision} * \text{Sensitivity} / (\text{Precision} + \text{Recall}) \quad (6)$$

RESULTS AND DISCUSSION

Results of BOW and SVM classifier

The results of BOW feature extraction and SVM classifier on the SPAM dataset used as seen in Table 1. The accuracy of 0.979 suggests that the model correctly classified 97.9% of the data points in the evaluation set. This indicates a high level of overall

correctness. A precision of 1.0 is a perfect score. It means that every instance the model predicted as positive was indeed a true positive. This indicates a high degree of confidence in the model's positive predictions. A recall of 0.85 means the model correctly identified 85% of the actual positive instances. This implies that 15% of the actual positive instances were missed. An F1 score of 0.917 suggests a good balance between the model's precision and recall. A high F1 score indicates that the model performs well in identifying positive instances while also maintaining a low rate of false positive predictions.

Table 1 BOW and SVM classifier

S/N	Metrics	Values
1	Accuracy	0.979
2	Precision	1.0
3	Recall	0.85
4	F1 score	0.917

Results of BOW and LR classifier

The results of BOW feature extraction and LR classifier on the SPAM dataset used as seen in Table 2. An accuracy of 0.978 suggests that the model correctly classified 97.9% of the data points in the evaluation set. This indicates a high level of overall correctness. A precision of 1.0 is a perfect score. It means that every instance the model predicted as positive was indeed a true positive. There were no false positive predictions. This indicates a high degree of confidence in the model's positive predictions. A recall of 0.84 means the model correctly identified 84% of the actual positive instances. This implies that 16% of the actual positive instances were missed. An F1 score of 0.913 suggests a good balance between the model's precision and recall. A high F1 score indicates that the model performs well in identifying positive instances while also maintaining a low rate of false positive predictions.

Table 2 BOW and LR classifier

S/N	Metrics	Values
1	Accuracy	0.978
2	Precision	1.0
3	Recall	0.84
4	F1 score	0.913

Results of BOW with PCA and SVM classifier

The results of BOW feature extraction with PCA and the SVM classifier on SPAM dataset used as seen in Table 3. An accuracy of 0.980 suggests that the model correctly classified 97.9% of the data points in the evaluation set. This indicates a high level of overall correctness. A precision of 1.0 is a perfect score. It means that every instance the model predicted as positive was indeed a true positive. There were no false positive predictions. This indicates a high degree of confidence in the model's positive predictions. A recall of 0.85 means the model correctly identified 85% of the actual positive instances. This implies that 15% of the actual positive instances were missed. An F1 score of 0.920 suggests a good balance between the model's precision and recall. A high F1 score indicates that the model performs well in identifying positive instances while also maintaining a low rate of false positive predictions

Table 3 SVM for BOW with PCA

S/N	Metrics	Values
1	Accuracy	0.980
2	Precision	1.0
3	Recall	0.853
4	F1 score	0.920

Results of BOW with PCA and LR classifier

The results of BOW feature extraction with PCA and LR classifier on the SPAM dataset used as seen in Table 4. The accuracy of 0.973 suggests that the model correctly classified 97.3% of the data points in the evaluation set. This indicates a high level of overall correctness. A precision of 0.96 is a good score. It means that every instance the model predicted as positive was indeed a true positive. This indicates a high degree of confidence in the model's positive predictions. A recall of 0.83 means the model correctly identified 83% of the actual positive instances. This implies that 17% of the actual positive instances were missed. An F1 score of 0.892 suggests a good balance between the model's precision and recall. A high F1 score indicates that the model performs well in identifying positive instances while also maintaining a low rate of false positive predictions

Table 4 BOW with PCA and LR classifier

S/N	Metrics	Values
1	Accuracy	0.973
2	Precision	0.962
3	Recall	0.833
4	F1 score	0.893

Result of TF-IDF and SVM classifier

The results of TF-IDF feature extraction and SVM classifier on the SPAM dataset used as seen in Table 5. The accuracy of 0.982 suggests that the model correctly classified 98.2% of the data points in the evaluation set. This indicates a high level of overall correctness. A precision of 1.0 is a perfect score. It means that every instance the model predicted as positive was indeed a true positive. This indicates a high degree of confidence in the model's positive predictions. A recall of 0.87 means the model correctly identified 87% of the actual positive instances. This implies that 13% of the actual positive instances were missed. An F1 score of 0.928 suggests a good balance between the model's precision and recall. A high F1 score indicates that the model performs well in identifying positive instances while also maintaining a low rate of false positive predictions

Table 5 TF-IDF and SVM classifier

S/N	Metrics	Values
1	Accuracy	0.982
2	Precision	1.0
3	Recall	0.867
4	F1 score	0.929

Result of TF-IDF and LR classifier

The results of TF-IDF feature extraction and LR classifier on the SPAM dataset used as seen in Table 6.The accuracy of 0.966 suggests that the model correctly classified 96.6% of the data points in the evaluation set. This indicates a high level of overall correctness. A precision of 0.99 is a good score. It means that every instance the model predicted as positive was indeed a true positive. This indicates a high degree of confidence in the model's positive predictions. A recall of 0.76 means the model correctly identified 76% of the actual positive instances. This implies that 24% of the actual positive instances were missed. An F1 score of 0.860 suggests a good balance between the model's precision and recall. A high F1 score indicates that the model performs well in identifying positive instances while also maintaining a low rate of false positive predictions

Table 6 TF-IDF and LR

S/N	Metrics	Values
1	Accuracy	0.967
2	Precision	0.99
3	Recall	0.76
4	F1 score	0.86

Result of TF-IDF with PCA and SVM classifier

The results of TF-IDF feature extraction with PCA and SVM classifier on the SPAM dataset used as seen in Table 7. The accuracy of 0.980 suggests that the model correctly classified 98.0% of the data points in the evaluation set. This indicates a high level of overall correctness. A precision of 0.97 is a good score. It means that every instance the model predicted as positive was indeed a true positive. This indicates a high degree of confidence in the model's positive predictions. A recall of 0.88 means the model correctly identified 88% of the actual positive instances. This implies that 12% of the actual positive instances were missed. An F1 score of 0.923 suggests a good balance between the model's precision and recall. A high F1 score indicates that the model performs well in identifying positive instances while also maintaining a low rate of false positive predictions

Table 7 TF-IDF with PCA and SVM

S/N	Metrics	Values
1	Accuracy	0.980
2	Precision	0.97
3	Recall	0.88
4	F1 score	0.923

Result of TF-IDF with PCA and LR classifier

The results of TF-IDF feature extraction with PCA and LR classifier on the SPAM dataset used as seen in Table 8 The accuracy of 0.945 suggests that the model correctly classified 94.5% of the data points in the evaluation set. This indicates a high level of overall correctness. A precision of 0.978 is a good score. It means that every instance the model predicted as positive was indeed a true positive. This indicates a high degree of confidence in the model's positive predictions. A recall of 0.60 means the model correctly identified 60 % of the actual positive instances. This implies that 40 % of the actual positive instances were missed. An F1 score of 0.748 suggests a good balance between the model's precision and recall. A high F1 score indicates that the model performs well in identifying positive instances while also maintaining a low rate of false positive predictions

Table 8 TF-IDF with PCA and LR classifier

S/N	Metrics	Values
1	Accuracy	0.945
2	Precision	0.97
3	Recall	0.61
4	F1 score	0.74

Result of N gram feature extraction and SVM

The results of N -gram feature extraction and SVM classifier on the SPAM dataset used as seen in Table 9. The accuracy of 0.976 suggests that the model correctly classified 97.6% of the data points in the evaluation set. This indicates a high level of overall correctness. A precision of 0.99 is a perfect score. It means that every instance the model predicted as positive was indeed a true positive. This indicates a high degree of confidence in the model's positive predictions. A recall of 0.83 means the model correctly identified 83% of the actual positive instances. This implies that 17% of the actual positive instances were missed. An F1 score of 0.905 suggests a good balance between the model's precision and recall. A high F1 score indicates that the model performs well in identifying positive instances while also maintaining a low rate of false positive predictions

Table 9 N gram feature extraction and SVM

S/N	Metrics	Values
1	Accuracy	0.977
2	Precision	0.99
3	Recall	0.83
4	F1 score	0.91

The result of N-gram FE and LR

The results of N-gram feature extraction and LR classifier on the SPAM dataset used as seen in Table 10 The accuracy of 0.956 suggests that the model correctly classified 95.6% of the data points in the evaluation set. This indicates a high level of overall correctness. A precision of 1.0 is a perfect score. It means that every instance the model predicted as positive was indeed a true positive. This indicates a high degree of confidence in the model's positive predictions. A recall of 0.67 means the model correctly identified 67% of the actual positive instances. This implies that 33% of the actual positive instances were missed. An F1 score of 0.804 suggests a good balance between the model's precision and recall. A high F1 score indicates that the model performs well in identifying positive instances while also maintaining a low rate of false positive predictions

Table 10 N gram feature extraction and LR

S/N	Metrics	Values
1	Accuracy	0.956
2	Precision	1.0
3	Recall	0.67
4	F1 score	0.80

Science World Journal Vol. 20(No 2) 2025 https://dx.doi.org/10.4314/swj.v20i2.20
www.scienceworldjournal.org
ISSN: 1597-6343 (Online), ISSN: 2756-391X (Print)
Published by Faculty of Science, Kaduna State University

Result N gram feature extraction with PCA and SVM
The results of N-gram feature extraction with PCA and SVM classifier on the SPAM dataset used as seen in Table 11. The accuracy of 0.978 suggests that the model correctly classified 97.8% of the data points in the evaluation set. This indicates a high level of overall correctness. A precision of 0.96 is a good score. It means that every instance the model predicted as positive was indeed a true positive. This indicates a high degree of confidence in the model's positive predictions. A recall of 0.87 means the model correctly identified 87% of the actual positive instances. This implies that 13% of the actual positive instances were missed. An F1 score of 0.916 suggests a good balance between the model's precision and recall. A high F1 score indicates that the model performs well in identifying positive instances while also maintaining a low rate of false positive predictions

Table 11 N gram feature extraction with PCA and SVM

S/N	Metrics	Values
1	Accuracy	0.978
2	Precision	0.963
3	Recall	0.87
4	F1 score	0.91

N gram feature extraction with PCA and LR
The results of N-gram feature extraction with PCA and LR classifier on the SPAM dataset used as seen in Table 12. The accuracy of 0.938 suggests that the model correctly classified 93.8% of the data points in the evaluation set. This indicates a high level of overall correctness. A precision of 0.97 is a good score. It means that every instance the model predicted as positive was indeed a true positive. This indicates a high degree of confidence in the model's positive predictions. A recall of 0.55 means the model correctly identified 85% of the actual positive instances. This implies that 45% of the actual positive instances were missed. An F1 score of 0.706 suggests a good balance between the model's precision and recall. A high F1 score indicates that the model performs well in identifying positive instances while also maintaining a low rate of false positive predictions

Table 12 N gram feature extraction with PCA and LR

S/N	Metrics	Values
1	Accuracy	0.938
2	Precision	0.976
3	Recall	0.55
4	F1 score	0.706

Result of BERT feature extraction and SVM
The results of BERT feature extraction and SVM classifier on the SPAM dataset used as seen in Table 13. The accuracy of 0.989 suggests that the model correctly classified 98.9% of the data points in the evaluation set. This indicates a high level of overall correctness. A precision of 0.97 is a good score. It means that every instance the model predicted as positive was indeed a true positive. This indicates a high degree of confidence in the model's positive predictions. A recall of 0.94 means the model correctly identified 94% of the actual positive instances. This implies that 6% of the actual positive instances were missed. An F1 score of 0.959 suggests a good balance between the model's precision and recall. A high F1 score indicates that the model performs well in identifying positive instances while also maintaining a low rate of false positive predictions

Table 13 BERT feature extraction and SVM

S/N	Metrics	Values
1	Accuracy	0.989
2	Precision	0.972
3	Recall	0.94
4	F1 score	0.95

Result of BERT feature extraction and LR
The results of BERT feature extraction and LR classifier on the SPAM dataset used as seen in Table 14. The accuracy of 0.991 suggests that the model correctly classified 99.1% of the data points in the evaluation set. This indicates a high level of overall correctness. A precision of 0.99 is a perfect score. It means that every instance the model predicted as positive was indeed a true positive. This indicates a high degree of confidence in the model's positive predictions. A recall of 0.99 means the model correctly identified 99% of the actual positive instances. This implies that 1% of the actual positive instances were missed. An F1 score of 0.990 suggests a good balance between the model's precision and recall. A high F1 score indicates that the model performs well in identifying positive instances while also maintaining a low rate of false positive predictions

Table 14 BERT feature extraction and LR

S/N	Metrics	Values
1	Accuracy	0.99
2	Precision	0.99
3	Recall	0.99
4	F1 score	0.99

The visualization results of the BERT feature extraction with PCA and LR classifier on SPAM dataset used as seen in Figure 2

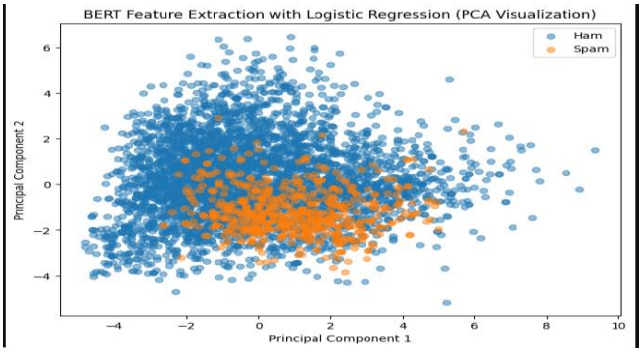

Figure 2 BERT feature extraction with PCA and LR

The visualization results of the TF-IDF feature extraction with PCA and LR classifier on SPAM dataset used as seen in Figure 3

Science World Journal Vol. 20(No 2) 2025 https://dx.doi.org/10.4314/swj.v20i2.20
www.scienceworldjournal.org
ISSN: 1597-6343 (Online), ISSN: 2756-391X (Print)
Published by Faculty of Science, Kaduna State University

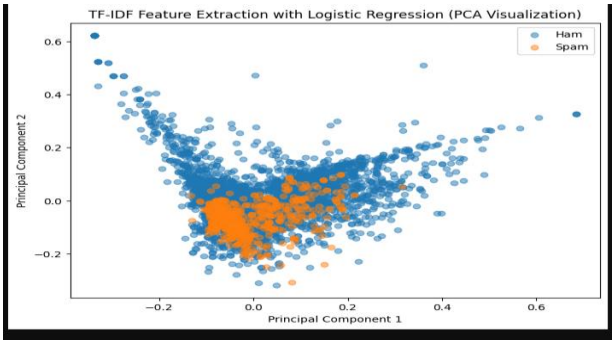

Figure 3 TF-IDF feature extraction with PCA and LR

The visualization results of the N-gram feature extraction with PCA and LR classifier on SPAM dataset used as seen in Figure 4

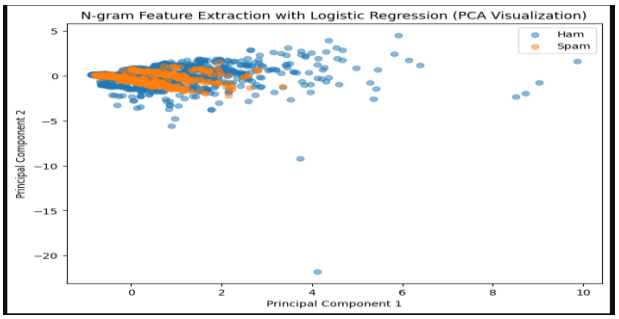

Figure 4 N-gram feature extraction with PCA and LR classifier

The visualization results of the BOW feature extraction with PCA and LR classifier on SPAM dataset used as seen in Figure 5

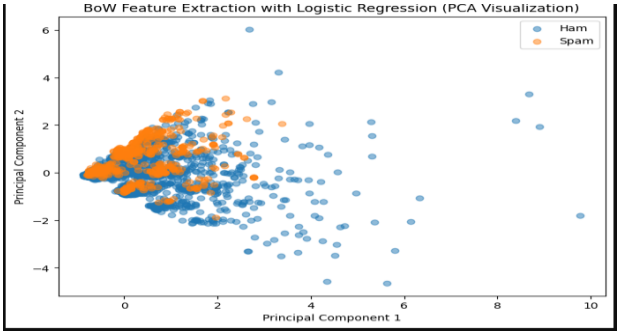

Figure 5 BOW feature extraction with PCA and LR classifier

DISCUSSION OF RESULTS

The Comparison of the performance of SVM and LR classifier with different feature extraction techniques on the SPAM dataset as shown in Table 15 and Figure 6. The results were used to justify the question on how does the five feature extraction techniques compare in terms of classification accuracy on the two classifiers.

Table 15 Comparison of FE techniques on SVM and LR classifier

FE	SVM Accuracy	LR Accuracy
BOW	0.979	0.978
BOW with PCA	0.980	0.973
TF -IDF	0.982	0.967
TF -IDF with PCA	0.980	0.945
N gram	0.977	0.956
N gram with PCA	0.978	0.938
BERT	0.989	0.99

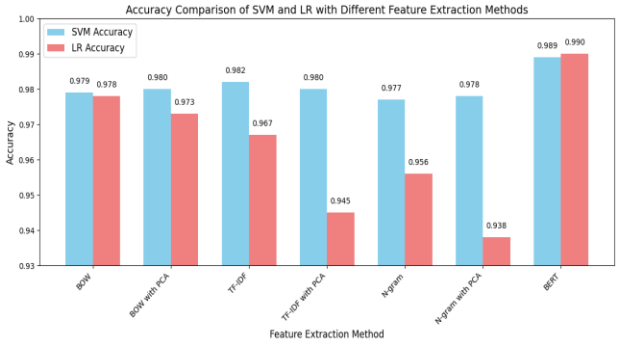

Figure 6 Comparison of SVM and LR

From the study carried out the BERTS FE usually lead to the highest accuracy for the experiment carried out, while both SVM and LR achieved their best accuracy of 0.989 and 0.990 respectively as seen in Table 15, it can be seen that the power of BERTS contextualized inserting in capturing rich semantic information

TF -IDF act well for SVM: TF-IDF features output the highest accuracy for SVM (0.982) among the non-BERT methods used. This pointed out that the TF-IDF adequately represented the document content of the SPAM by examining the words based on their relevance.

The PCA influence changes: When PCA is used for dimensionality reduction, it has mixed effects on the accuracy of the model especially in datasets like SPAM. It can be seen that in some cases, it slightly improves the accuracy for example SVM with BOW and TF-IDF, while in the other instance of LR with TF-IDF and Ngram it leads to a decrease in their accuracy. This suggests that Models effectiveness depends on the specific dataset and feature extraction methods employed which in line with the study of.(Santoso & Priyadi 2025).

In order to justify the question, are there significant differences in performance between the two classifiers when using the different feature extraction method?

BOW and N-gram FE techniques, these two give a brilliant performance. It is observed that BOW and N-gram FE achieved reasonable accuracy for both SVM and LR, but generally lower than TF-IDF and BERT. This indicates that these techniques capture some information but the performance cannot be compared with advanced FE techniques like BERT.

BOW versus BOW with PCA: It is observed that when PCA is combined with BOW it slightly improves SVM accuracy but slightly reduced LR accuracy. This pointed out that PCA might help in reducing noise and redundancy in BOW features for SVM, but might lose some critical information in LR as seen in Figure 7. The result shows significant differences in the performance of the classifier when use different FE

Science World Journal Vol. 20(No 2) 2025 https://dx.doi.org/10.4314/swj.v20i2.20
www.scienceworldjournal.org
ISSN: 1597-6343 (Online), ISSN: 2756-391X (Print)
Published by Faculty of Science, Kaduna State University

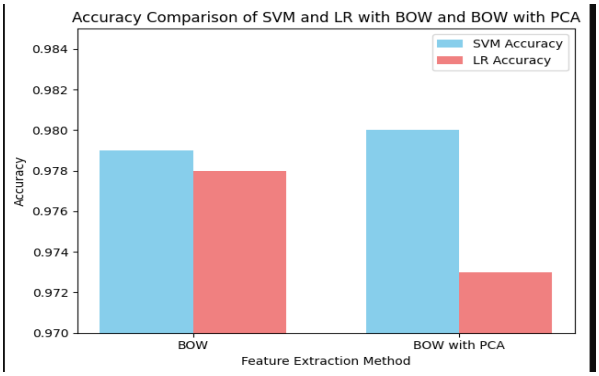

Figure 7 Accuracy of BOW versus BOW with PCA

In order to answered the question are there any computational trade-offs associated with the different feature extraction methods? TF-IDF versus TF-IDF with PCA: the study revealed that PCA has a more pronounced negative impact on LR accuracy when applied to TF-IDF features. This can be as a result of PCA removing some critical information captured by TF-IDF which is beneficial for LR'S performance as seen in Figure 8

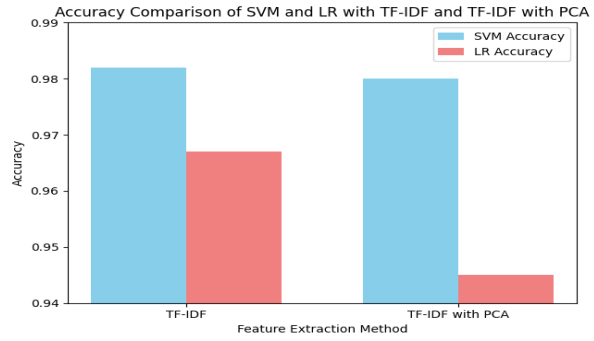

Figure 8 TF-IDF versus TF-IDF with PCA

N-Gram versus N-gram with PCA: the simulation carried out in the study revealed that the performance of PCA with -gram is similar to the result of TF-IDF with PCA. PCA reduces LR accuracy when applied to N gram features. This inform the observation that PCA's effectiveness depends on the interaction between the FE methods and classification algorithms employed as seen in Figure 9.

O-

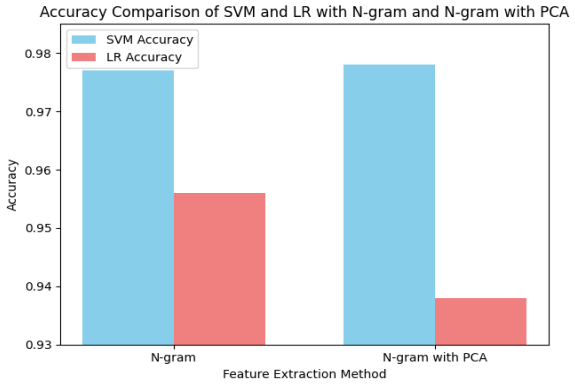

Figure 9 N-gram versus N-gram with PCA

BERT versus BERT with PCA: BERTS shows a high accuracy which suggests that its features are already quite effective and applying PCA might not be necessary or even beneficial. The study is compared with the finding of Fadhel et al (2025) with TF-IDF accuracy of 94%, Mohtasham et al. (2024) accuracy of 89% and Wamidh k. Mutlag et al. (2022) accuracy of 98%. The study outperformed the existing models in the reviewed literature in term of the accuracy of the TF-IDF accuracy of 98% for SVM and 99% for BERT with the LR model developed.

Conclusion

The study concluded that Models effectiveness depends on the specific dataset and feature extraction methods employed so also, the accuracy results highlight the importance of choosing appropriate feature extraction techniques and considering the effect of dimensionality reduction like PCA, BERT features usually yield the highest accuracy followed by TF-IDF, while BOW and N-gram provide decent performance on the SPAM dataset used. The effect of PCA varies depending on the specific dataset and FE method. The results of this finding will guide research in this area in the selection of feature extraction and dimensionality reduction strategies for developing good classification Models. The study recommends that careful selection of FE methods will optimize the model performance and achieve a higher accuracy. Further works can be done with different dataset, use of different FE techniques and different Deep learning algorithm.

Conflict of Interest: The Authors have no conflict of interest
Declaration: We declare that the authors carried out the original work.

REFERENCES

Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to Information Retrieval. Cambridge University Press.

Sebastiani, F. (2002). Machine learning in automated text categorization. ACM computing surveys (CSUR), 34(1), 1-47.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and Trends® in Information Retrieval, 2(1–2), 1-135.

Ogunsanwo Gbenga. O. (2025). A Comparative Analysis of Ensemble Machine Learning Algorithms for Bank Customer Churn Prediction, University of Ibadan Journal of Science and Logics in ICT Research

(UIJSLICTR). Vol. 13, No.1, pp. 80 –94.

Omotunde Ayokunle A., Ogunsanwo, Gbenga, O., Adekola, Olubukola, zang Aaron, and Abel Samuel B (2022) COMPARATIVE ANALYSIS OF DATA MINING TECHNIQUES FOR MOVIE PREDICTION FUDMA Journal of Sciences (FJS) Vol. 6 No. 6, December, 2022, pp 224 - 228 224

Ogunsanwo, G. O., Odulaja, P. T., Omotunde, A. A. & Solanke, O. O. (2025). Air quality index prediction using deep learning for Lagos State in Nigeria. Lafia Journal of Scientific and Industrial Research, 3(1), 108 – 117. https://doi.org/10.62050/ljsir2025.v3n1.450

Salton, G., & McGill, M. J. (1986). Introduction to modern information retrieval. McGraw-Hill, Inc.

Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. Journal of documentation, 28(1), 11-21.

Santoso, L., & Priyadi, P. (2025). Comparative Study of Feature Engineering Techniques for Predictive Data Analytics. Journal of Technology Informatics and Engineering. February 2025.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems (pp. 3111-3119).

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 1532-1543)

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) (pp. 4171-4186).

Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine learning, 20(3), 273-297

Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

Fernández-Delgado, M., Sánchez-Chacón, E., Bielza, C., & Larrañaga, P. (2014). Do we need hundreds of classifiers to solve real world classification problems?. Journal of Machine Learning Research, 15(1), 3145-3181.

Zahraa Fadhel, Hussien Attia & Yossra Hussain Ali A Comparative Analysis of Feature Extraction Techniques for Fake Reviews Detection Fusion: Practice and Applications (FPA) Vol. 17, No. 02. PP. 161-172, 161 DOI: https://doi.org/10.54216/FPA.170212

Farideh Mohtasham, MohamadAmin Pourhoseingholi , Seyed Saeed Hashemi Nazari , Kaveh Kavousi & Mohammad Reza Zali1 (2025) Comparative analysis of feature selection techniques for COVID-19 dataset Scientifc Reports |https://doi.org/10.1038/s41598-024-69209-6

Hemdanou Abderrafik Laakel , Mohammed Lamarti Sefian, Youssef Achtoun & Ismail Tahiri (2024) Comparative analysis of feature selection and extraction methods for student performance prediction across different machine learning models www.sciencedirect.com/journal/computers-and-education-artificial- intelligenceorg/10.1016/j.caeai. .100301

Zhao Shuai1, Diao Xiaolin1, Yuan Jing , Huo Yanni , Cui Meng , Wang Yuxin and Zhao Wei(2022)Comparison of diferent feature extraction methods for applicable automated ICD coding BMC Medical Informatics and Decision Making (2022) 22:11 https://doi.org/10.1186/s12911-022-01753-5

Wamidh K. Mutlag, Shaker K. Ali, Zahoor M. Aydam and Bahaa H.Taher (2020), Feature Extraction Methods: A Review doi:10.1088/1742- 6596/1591/1/012028

Federico Calesella , Alberto Testolin, Michele De Filippo De Grazia and Marco Zorzi(2021) A comparison of feature extraction methods for prediction of neuropsychological scores from functional connectivity data of stroke patients Brain Informatics https://doi.org/10.1186/s40708-021-00129-1

Zhang, W., Yoshida, T., & Tang, X. (2010). A comparative study of TFIDF, LSI and multi-words for text classification. Expert Systems with Applications, 38(3), 2758-2765.

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: a review and recent developments. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374(2065), 20150202.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to information retrieval. Cambridge university press.

Cavnar, W. B., & Trenkle, J. M. (1994). N-gram-based text categorization. Ann Arbor MI, 48113(2), 161-175.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.