ISSN 1330-3651 (Print), ISSN 1848-6339 (Online) https://doi.org/10.17559/TV-20240917001995
Original scientific paper

# Investigating the Contribution of Structural and Contextual Features for Spam Detection

Ulin NUHA, Chih-Hsueh LIN*

**Abstract:** One of the detrimental consequences resulting from the rapid and effortless dissemination of information is the high flow of spam messages in user's electronic devices. Various studies have attempted to conduct spam detection by proposing several approaches for spam detection based on input features. However, the effectiveness and efficiency of spam detection need improvement. Therefore, this paper investigates significant feature schemes for spam detection, including structural features and contextual representation. We propose a hybrid approach combining both features and evaluate it using various machine learning algorithms for short message service (SMS) spam classification. Experimental results show that relying solely on contextual representation outperforms structural features, achieving an accuracy of 92.56%. However, the hybrid approach, combining both structural and contextual features, achieves superior results with an accuracy of 93.22% and an *F*-score of 95.12%. Among the classifiers tested, random forest demonstrated the most robust performance, consistently achieving accuracy above 90% across all feature extraction schemes. The findings highlight the potential of combining structural and contextual features to enhance spam detection performance, with practical implications for telecommunication providers aiming to improve SMS filtering accuracy.

**Keywords:** classifier algorithms; contextual representation; feature analysis; hybrid scheme; spam detection

## 1 INTRODUCTION

Due to the rapid advancement of technology, sending messages to multiple recipients can be accomplished in a short time. However, many of these messages are unwanted, such as scams and spam messages. In November 2022, approximately one billion spam messages were sent daily [1]. Spam messages often take the form of unreasonable gifts, debt offers, or bill payments from unknown senders [2, 3]. The structure of spam messages typically differs from that of normal messages. Spam messages frequently contain untidy patterns, such as a high number of uppercase letters, unusual punctuation, links, and spam-related words [4, 5]. As spammers become increasingly sophisticated in their techniques, traditional approaches to detecting spam have proven inadequate, necessitating the development of advanced spam detection techniques [6]. Spammers can manipulate spam messages through synonym replacement techniques, non-spam word injections, and other methods that spam filters fail to identify. Thus, detecting and filtering spam messages are critical tasks.

The study of spam detection and prevention has attracted significant attention from both academia and industry sectors. In recent years, various machine learning and data mining algorithms have been introduced to enhance spam detection capabilities. These techniques exploit the power of intelligent algorithms to analyze message content, identify patterns, and distinguish between legitimate and non-legitimate messages. The existing studies of spam text detection or classification generally focus on two approaches, i.e., structural-based features and contextual-based representation. Various studies have conducted spam detection utilizing structural-based features as the model input. These structural features come from attributes in the message text, including the number of non-alphanumeric characters, capital letters, words per sentence, and others [7, 8]. Such attributes become the input for machine learning classifiers, enabling them to learn and recognize spam patterns effectively. The other approach learns the spam text pattern based on contextual representation. In this approach, the spam detection model utilizes contextual word representations of spam messages in embedding vectors as the input for machine learning [9, 10]. Compared to email spam, spam messages on mobile phones are characterized by shorter text content [9]. These messages are often more challenging to analyze due to the frequent use of slang and abbreviations [7]. Moreover, the limited availability of datasets for training poses a significant challenge for classification tasks of short message services (SMS) texts using machine learning [11].

Another significant challenge in SMS spam detection lies in the brevity and informal nature of messages, which often contain slang or abbreviations [12]. This paper focuses on investigating the most effective feature scheme for machine learning models to accurately learn and detect spam message patterns, particularly in SMS. This paper analyzes the significant features indicating spam messages based on structural features, contextual representation features, and the combination of these two features. Structural features are derived from attributes that are presented in spam message texts. While, contextual representation features are obtained by exploiting bidirectional encoder representations from transformers (BERT). Then, this study leverages a hybrid feature extraction scheme that combines structural and contextual representation features for spam SMS detection. To the best of our knowledge, this is the first work to simultaneously utilize both types of features for this purpose. The study evaluates the effectiveness of structural features and contextual representations in identifying spam messages. Once the features are extracted, several machine learning algorithms, including support vector classifier (SVC), multi-layer perceptrons (MLP), random forest, and others, are employed to classify spam messages. Overall, the contribution of our work can be summarized below.

1. This study investigates the performance of two types of extracted features from message texts for spam detection, namely, the structural features and the contextual representation.
2. This study employs various machine learning schemes for spam classification tasks to reveal the best classifier in this study.

3. This paper also proposes a hybrid feature extraction scheme combining structural features and contextual representation to better capture message information pattern for spam detection.

4. Our results reveal that the fusion of structural features and contextual representations provides the most effective input scheme for spam detection models.

The rest of this paper is organized as follows. We provide a comprehensive review of related works in section 2. Then, we introduce the proposed method and its implementation in section 3. We present the description of the used dataset and the experimental details in section 4. We present the obtained experimental results and their analysis in Section 5. We highlight the conclusions and present a brief overview of future works in section 6.

## 2 RELATED WORK

Most spam detection schemes are focused on analyzing content-based and non-content-based features [13]. Content-based feature schemes emphasize the textual features and attributes within the message texts. This study conducts spam detection based on the content-based approach, utilizing several features. Previous studies have explored spam detection using the content-based approach, particularly by exploiting structural features embedded in the text. This approach can be performed since spam messages have several distinct attributes, such as an uppercase alphabetic presence, use of uniform resource locators (URL), word frequencies, and others [8, 14]. These attributes serve as valuable model inputs for spam detection. For instance, Ghosh and Senthilrajan utilized such attributes for the model input of various machine learning techniques to detect spam messages, with the random forest algorithm demonstrating the best performance [15]. Faris et al. proposed spam identification approach based on the most significant features in the dataset using a genetic algorithm (GA) and random weight network (RWN) [16]. They selected about 25 features obtained in the structural attributes of the dataset, such as the number of uppercase characters, spam-related words, and others. Another study presented in [17] exploited the naïve Bayesian (NB) classifier as the classifier model for Arabic spam detection. The work extracted features from SMS text attributes, including the number of words and the presence of URLs. Its experimental results indicated that the accuracy of spam SMS detection is good but not exceptional. Batra et al. proposed spam classification in email messages utilizing *k*-nearest neighbours (*k*-NN) integrated with various optimization schemes [7]. They considered several features from structural attributes such as word frequency, character frequency, and capital letters as key features. Their main findings were that the use of the Manhattan distance for *k*-NN yielded the most satisfactory distance metric for spam classification.

Various studies have also conducted spam detection based on linguistic features by exploiting textual or contextual representations of spam messages. For instance, Rojas-Galeano developed an SMS spam detection model focusing on word representation using several techniques, including bag-of-words (BoW), term frequency-inverse document frequency (TFIDF), and BERT [18]. After obtaining word representation vectors, several machine learning classifiers were applied to perform spam classification. The results showed that BERT encoding achieved the best score in the evaluation metrics. In another study, a comprehensive SMS spam detection called SpotSpam was developed to address the challenge of spam detection on a smaller number of message characters using textual features [9]. SpotSpam categorized spam data into multiple classes based on the sender's intention and utilized BERT and its variants to learn linguistic features from spam messages. The experimental results revealed that a combination of DistilBERT and support vector machine (SVM) achieved high accuracy. Similarly, another study tackled spam detection in both SMS and Twitter messages by employing BERT to extract linguistic features [19]. Compared to the other feature extraction methods, the BERT model consistently demonstrated superior performance, highlighting the potential of contextual representations for spam detection.

Most of the aforementioned works conducted spam detection by relying on either structural features or linguistic features. However, some studies have explored the integration of multiple feature types to enhance spam detection performance. For instance, Andresini et al. introduced a comprehensive study on spam filtering that considered various features [20]. Their review spam detection model employed multi-view features, combining content-based and behavior-based attributes. The content-based features were obtained from the linguistic context in the spam texts, while the behaviour-based features came from the behaviour of the user's reviews, such as the number of given reviews, the content similarity, and others. Although effective for detecting review spam, this approach is not applicable to SMS spam detection due to the lack of user behavior-related information in SMS datasets. This highlights the need for specialized methods that focus on features inherent to SMS texts, such as structural and linguistic representations. Therefore, our study investigates the impact of multi-view features for the model input to determine spam messages considering structural features and linguistic context features, simultaneously or separately. To the best of our knowledge, this research is the first to explore multi-view features in the detection of spam messages based on the structural features and the contextual representation feature. After extracting these features, we utilize several machine learning classifiers to evaluate whether the incorporation of multi-view aspects enhances the performance of spam message detection.

## 3 METHODS

Our proposed spam detection method exploits multiple features from the message text. The proposed method generally comprises two main phases: Phase 1involves analyzing and extracting both structural features and contextual word representation features from the message text. Phase 2 utilizes the extracted features from Phase 1 to perform the classification of messages using various machine learning algorithms.

### 3.1 Structural Features

Spam detection can be categorized as a conventional text classification procedure. However, the text structure of spam messages differs significantly from that of normal messages. Spam messages frequently contain salient attributes, including non-alphanumeric characters such as "!" or "#", URLs, and numeric characters such as dates or phone numbers [17, 21]. Additionally, another significant structural feature in spam text is the presence of typical spam keywords [22]. These spam keywords consist of recurring groups of identical words. To determine the spam keywords, we explore the most frequent words appearing in labelled spam messages, taking the most frequently appearing words among 30 words in the labelled spam message. The structural features became critical attributes for spam detection, as they focus on the message's structure without relying on an understanding of its linguistic context. Through these features, we identify patterns within the structure of the suspicious message. Then, we select seven attributes as the structural features based on previous studies. Tab. 1 summarizes the most prominent structural features, with citations included based on findings from previous studies.

**Table 1** The description of structural features

| No | Feature name | Description |
|---|---|---|
| 1 | Num_word [5] | Number of words in a message. |
| 2 | Num_char [5] | Number of characters in a message. |
| 3 | Uppercase_char [16] | Number of uppercase characters in a message. |
| 4 | Non_alphanumeric_ char [17, 21] | Number of non-alphanumeric characters in a message. |
| 5 | Numeric_char [17] | Number of numeric characters found in a message. |
| 6 | Presence_URL [8, 21] | Presence of URLs in a message. |
| 7 | Spam_words [22] | Number of appeared spam words. |

### 3.2 Feature Importance

After determining several attributes in the structural features, we can calculate the feature importance of each attribute. Feature importance refers to identifying the features exhibiting strong dependencies to effectively convey the forecasting results, evaluated using a random forest algorithm [23]. Evaluations based on the importance feature are often calculated using common metrics such as the Gini index [24], which assesses the capability of a feature to segregate samples from distinct categories. The algorithm should attain all features and discover a value to discriminate distinct categories and decrease the Gini index. For example, assume a set of $n$ features in $F = \{f_1, f_2, f_3, \ldots, f_n\}$. If the sample set in node $s$ of a decision tree has $C$ categories, the formula to calculate Gini index $GI$ is defined as.

$$GI_s = \sum_{i=1}^{C} p_i(1-p_i) = 1 - \sum_{i=1}^{C} p_i^2 \quad (1)$$

where $p_i$ indicates the proportion of observed samples in the $i$-th category in node $s$. The feature importance of $f_j$ at node $s$ is the value change of $GI$ before and after branching in the decision tree:

$$VI_{js} = GI_s - GI_l - GI_r \quad (2)$$

where $VI_{js}$ is the variable importance measure of each feature ($f_j$). $GI_l$ and $GI_r$ refer to the Gini index values of new nodes after branching in the left and right. For feature $f_j$ in a specific node that belongs to set $S$, the feature importance of $f_j$ in its decision tree ($d$) is formulated as follows:

$$VI_{dj} = \sum_{s \in S} VI_{js} \quad (3)$$

### 3.3 Contextual Representation Feature

Since spam message detection can be categorized as a text classification, various studies have proposed spam detection based on textual or contextual representation using feature embeddings such as BoW, TFIDF, and others. In [18], a study was proposed to detect spam messages using various word representation models with experimental results in which the BERT model preserved excellent performance compared to the BoW and TFIDF encoders. The architecture of BERT comprises multiple blocks of transformer-encoders. Each encoder block comprises stacks of identical layers, including multi-head attention mechanisms and feed-forward networks, as shown in Fig. 1. In the figure, $E_i$ refers to the input embedding at the $i$-th input token, while $T_i$ denotes the final hidden vector of the input, and $T_r$ is the transformer-encoder block.

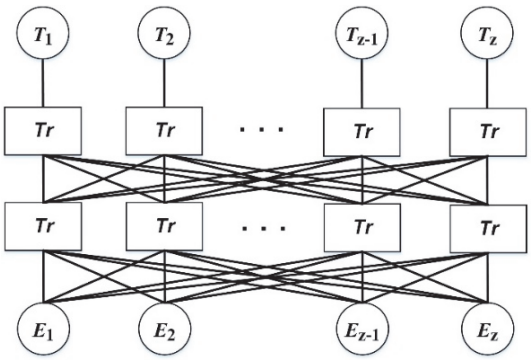


**Figure 1** The BERT architecture

The BERT model can achieve excellent performance in representing contextual features due to its pre-training on a large text corpus using a multi-task learning approach. This approach combines two objectives: masked language modeling (MLM) and next-sentence prediction (NSP), which enhance the model's ability to derive detailed sentence representations [25]. During the MLM phase in the pre-training process, the tokens are subjected to different treatments, with 15% of the token data randomly masked [26], 80% of the randomly masked data replaced with a (mask) token, 10% substituted with other tokens, and 10% left unchanged. This encourages the model to predict the masked tokens based on their context, thereby learning robust representations. In the NSP phase, the model is trained to determine the relationship between two sentences. Half of the pairs retain their original order and are labelled "Is Next", indicating a related relationship, while the remaining has their second sentence randomly chosen from the corpus and labelled "Not Next", indicating an unrelated relationship. After the pre-training process for

BERT is completed, we fine-tune the model for various downstream tasks, such as text classification.

The embedding stage of the BERT input representation is illustrated in Fig. 2. This stage incorporates three types of embeddings: token embedding, segment embedding, and position embedding. The input tokens in the BERT model are obtained after applying the WordPiece tokenization scheme. The segment embedding layer assigns every token to its respective input segment, after which the position embedding layer denotes the position of each token. Then, all embeddings are fused using element wise-operation.

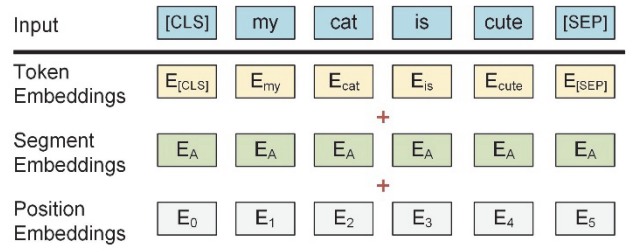


**Figure 2** The BERT input representation

### 3.4 Input Feature Scheme

As described earlier, this study utilizes structural and contextual representation features for the model input. Thus, we propose three schemes for the input feature, as illustrated in Fig. 3. For scheme 1, we only exploit the structural features as the model input for spam detection, as listed in Tab. 1. Scheme 2 leverages the contextual representation features of the message texts extracted the BERT model, as the model input. Scheme 3 utilizes the combination of the structural and contextual representation features as the model input. We combine both features to leverage their complementary strengths, enabling a more comprehensive understanding of spam message patterns.

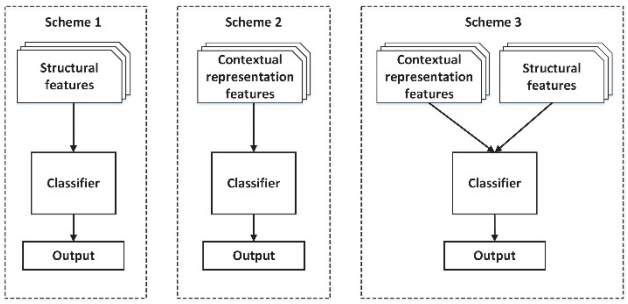


**Figure 3** Input feature scheme for spam detection

Structural features capture specific patterns, such as the presence of URLs, numeric characters, or uppercase characters, which are often indicative of spam messages, while contextual features extracted provide the semantic and linguistic meaning of the text. By combining these two feature types, the hybrid approach may obtain a more comprehensive representation of the data, allowing the model to detect spam messages that might be missed if only one feature type is used. This synergy enhances the model's ability to identify both structural anomalies and semantic inconsistencies, leading to enhance performance. These features in all schemes become the input for the classifiers to perform spam classification tasks. To evaluate the effectiveness of the proposed schemes, we apply various machine learning algorithms as classifiers, detailed in Tab. 3.

To implement the hybrid approach, we apply concatenation of the vectors of the structural and contextual representation features. The structural feature vector has a size of 7, as shown in Tab. 1, while the BERT model generates a contextual feature vector of size 768, as we utilize the base version of the BERT model. After concatenating these vectors, we apply the StandardScaler (https://scikit-learn.org/1.5/modules/generated/sklearn.preprocessing.StandardScaler.html) to normalize the combined vector before inputting it into the classifier. This scaling ensures that all features contribute proportionally to the classification task, avoiding potential bias from features with larger scales.

### 3.5 Evaluation Metrics

We propose several evaluation metrics to assess the performance of the spam detection models. Since the amount of class data is imbalanced, we exploit the balanced accuracy, precision, recall, and *F*-score. The balanced accuracy is a specialized accuracy metric designed to handle imbalanced data [27]. The precision indicates how the model produces more pertinent results than to irrelevant ones, whereas the recall signifies how the model retrieves a significant portion of the relevant results. The *F*-score is the harmonic mean of the precision and recall. The formulas for the evaluation metrics are as follows:

$$balanced\_accuracy = \frac{1}{2}\left(\frac{TP}{TP+FN}+\frac{TN}{TN+FP}\right) \tag{4}$$

$$precision = \frac{TP}{TP+FP} \tag{5}$$

$$recall = \frac{TP}{TP+FN} \tag{6}$$

$$F\text{-score} = \frac{2\times precision\times recall}{precision+recall} \tag{7}$$

A true positive (*TP*) occurs when both the predicted and actual values are positive. A false positive (*FP*) is the condition when the predicted value is positive, however the actual value is negative. Then, a false negative (*FN*) is the condition when the predicted value is negative, however the actual value is positive. Lastly, a true negative (*TN*) occurs when both the predicted and actual values are negative.

## 4 EXPERIMENTAL SETUPS

### 4.1 Dataset Introduction

For the experimental stage, we utilize a dataset of SMS texts in the Indonesian language, Indonesia has a high prevalence of SMS spam attacks. Based on a report from Truecaller [28], in 2021, Indonesia was the most vulnerable country to SMS spam attacks in Asia. In our study, the SMS text dataset is collected from [29]. We divide the text

messages into two classes of spam and non-spam messages. The total number of messages is 1143; with 574 spam messages and 569 non-spam messages. To assess robustness, we also include an additional dataset from Tandra et al. [30], which consists of 2535 spam messages and 811 normal messages. Then, we remove duplicate messages from the combined datasets. The spam messages encompass SMS texts that serve the purpose of commercial advertising, as well as those with illicit material or objectives, such as fraudulent content, fake news, and similar endeavors. We also divide the dataset into training data and testing data with a ratio of 0.8:0.2. Tab. 2 shows the samples of spam and non-spam SMS texts in the Indonesian language from our used dataset.

**Table 2** Spam and non-spam message samples

| SMS text | Description |
|---|---|
| GT-SHOP Disc.41% BB Z10 3jt BB Torch 2jt BB Davis 1jt Minat? Galaxy S4 3,5jt Cp.08994352577 Pin.25A1785B Gili-SMS versi Demo. Http://www/yusiwa.com | Spam |
| Disc.25% Domino's Pizza Value Range s/d 14 Agustus.Kunjungi store terdekat/hubungi 1500366/visit http://bit.ly/294KoZ6 Kode:VR25D | Spam |
| Maaf jika ada janji yang belum terpenuhi, jika ada janji boleh mengingatkan saya. | Non-spam |
| Ada mba, harganya 1,6jt/bln. Fasilitasnya air hangat, kamar mandi dalam, tv, tv kabel, wifi, meja, tempat tdr, dan lemari. | Non-spam |

### 4.2 Experimental Details

During the experimental stage, we utilized the BERT model to extract contextual representation features, employing SentenceTransformers [31] and IndoBERT [32] as the pre-trained model. For the classifiers, we exploit the random forest, *k*-NN, MLP, Gaussian NB (Naive Bayes), and others from Scikit-learn, a machine learning library for the Python programming language. The model parameters of these machine learning algorithms are shown in Tab. 3. The remaining parameter values were set to their default configurations as defined by the Scikit-learn library. We also perform *k*-fold cross-validation during the training process, to minimize overfitting and underfitting. *K*-fold cross-validation is a resampling scheme that divides the data into diverse subsets, with each subset in each iteration alternately becoming validation data [33]. Here, *k* represents the number of groups into which a given data sample is to be divided. We set the value of *k* as equal to five, indicating a five-fold cross-validation.

**Table 3** The description of structural features

| Classifier model | Parameters |
|---|---|
| Random forest | n_estimators = 100 |
| *K*-nearest neighbours | n_neighbors = 15 |
| Logistic regression | Penalty = 'l2', max_iter = 100 |
| Linear SVC | default parameters |
| SVC | Kernel = 'rbf', gamma = 0.01, *C* = 100 |
| MLP | hidden layer sizes = 10, alpha = 1, max iter = 1000 |
| Decision tree | max depth = 10, criterion = 'gini' |
| Gaussian NB | var_smoothing = 1e-9 |

## 5 RESULTS AND DISCUSSION

### 5.1 Importance of the Structural Features

As mentioned above, we utilize structural features in message texts for spam detection. To assess their contributions to spam classification, we evaluate the feature importance of each attribute using the random forest algorithm. The result of the feature importance evaluation is presented in Tab. 4. The "spam_words" feature achieves the highest feature importance value. This suggests that messages containing spam keywords strongly correlate with a malicious objective, consistent with our finding that more than 90% of spam messages contain at least one spam keyword. The second-most dominant feature is the "num_char" feature, which reflects the number of characters in a message. In our dataset, spam messages contain an average of approximately 123 characters, whereas non-spam messages have an average of around 65 characters. This significant difference highlights the potential of this feature to aid in distinguishing spam from non-spam messages. Conversely, the structural feature with the lowest feature importance is "presence_URL", reflecting that not all spam messages include URLs. Similarly, "non_alphanumeric _char" and the "numeric_char" features also show insignificant feature importance values. Thus, spam messages and non-spam messages do not have a prominent difference in the number of non alphanumeric and non numerical characters.

**Table 4** Feature importance results

| No | Feature name | Score |
|---|---|---|
| 1 | Num_word | 0.109 |
| 2 | Num_char | 0.228 |
| 3 | Uppercase_char | 0.193 |
| 4 | Non_alphanumeric_ char | 0.095 |
| 5 | Numeric_char | 0.097 |
| 6 | Presence_URL | 0.019 |
| 7 | Spam_words | 0.258 |

### 5.2 Comparison Results

Here, we present a comparative analysis of spam detection performance across three input schemes: structural features, contextual representation features, and the hybrid approach. We examine all schemes by exploiting various machine learning techniques to perform classification tasks. Then, we utilize several evaluation metrics, such as the accuracy, precision, recall, and *F*-score, to assess the performance. We perform the training process based on the *k*-fold cross-validation scheme, as mentioned before (that is, the training data and validation data subsets are split *k* times. Tab. 5 shows the result comparison using all structural features, while Tab. 6 and Tab. 7 exhibit the experimental results using the contextual representation feature and the hybrid scheme, respectively. All values in the evaluation metrics are expressed in the form of percentages.

**Table 5** Experimental results using all structural features

| Classifier | *Accuracy* | *Precision* | *Recall* | *F*-Score |
|---|---|---|---|---|
| Random forest | 91.86 | 95.12 | 95.09 | 95.00 |
| *K*-nearest neighbours | 89.50 | 93.69 | 93.65 | 93.48 |
| Logistic regression | 88.17 | 91.88 | 92.00 | 91.91 |
| Linear SVC | 88.75 | 91.72 | 91.74 | 91.73 |
| SVC | 88.61 | 92.57 | 92.68 | 92.54 |
| MLP | 88.29 | 92.47 | 92.59 | 92.44 |
| Decision tree | 91.34 | 93.89 | 93.92 | 93.88 |
| Gaussian NB | 88.69 | 91.17 | 91.00 | 91.07 |
| Average | 89.40 | 92.81 | 92.83 | 92.76 |

**Table 6** Experimental results using the contextual representation feature

| Classifier | *Accuracy* | *Precision* | *Recall* | *F*-Score |
|---|---|---|---|---|
| Random forest | 93.81 | 95.88 | 95.91 | 95.88 |
| *K*-nearest neighbours | 90.04 | 94.70 | 94.62 | 94.46 |
| Logistic regression | 95.01 | 96.47 | 96.47 | 96.47 |
| Linear SVC | 93.54 | 95.22 | 95.15 | 95.16 |
| SVC | 96.00 | 96.81 | 96.71 | 96.72 |
| MLP | 95.06 | 96.64 | 96.62 | 96.62 |
| Decision tree | 86.27 | 90.23 | 90.24 | 90.17 |
| Gaussian NB | 90.76 | 91.54 | 90.12 | 90.44 |
| Average | 92.56 | 94.69 | 94.48 | 94.49 |

**Table 7** Experimental results using hybrid scheme

| Classifier | *Accuracy* | *Precision* | *Recall* | *F*-Score |
|---|---|---|---|---|
| Random forest | 94.09 | 96.22 | 96.24 | 96.21 |
| *K*-nearest neighbours | 90.08 | 94.69 | 94.62 | 94.46 |
| Logistic regression | 95.61 | 96.95 | 96.95 | 96.94 |
| Linear SVC | 94.61 | 96.18 | 96.18 | 96.17 |
| SVC | 95.56 | 96.81 | 96.79 | 96.79 |
| MLP | 95.37 | 96.60 | 96.53 | 96.54 |
| Decision tree | 89.24 | 92.61 | 92.68 | 92.63 |
| Gaussian NB | 91.19 | 91.96 | 90.97 | 91.21 |
| Average | 93.22 | 95.25 | 95.12 | 95.12 |

The experimental results of spam detection using the structural features reveal that the random forest model has the highest values for the accuracy, precision, recall, and *F*-score at 91.86, 95.12, 95.09, and 95.00, respectively, followed by the decision tree model as the second-best performer. For spam detection using the contextual representation features, our experiment utilizes the BERT model to obtain the feature embedding. The results show that SVC achieves the best results for accuracy, precision, recall, and *F*-score at 96.00, 96.81, 96.71, and 96.72, respectively. For the experimental results of the hybrid scheme, logistic regression emerges as the best-performing model, achieving the highest accuracy, precision, recall, and *F*-score at 95.61%, 96.81%, 96.79%, and 96.17%, respectively. Then, the random forest model demonstrates consistent and stable performance across all input feature schemes, with evaluation metric scores exceeding 90.00. In three schemes of the model input, random forest achieves the highest value of accuracy average with the value of 93.25. Tab. 8 provides a detailed comparison of the performance of random forest as the classifier under the three different input feature schemes for spam classification. The hybrid approach proves to be the most effective scheme for spam detection when paired with the random forest classifier.

**Table 8** Experimental results using hybrid scheme

| Classifier | *Accuracy* | *Precision* | *Recall* | *F*-Score |
|---|---|---|---|---|
| Structural feature | 91.86 | 95.12 | 95.09 | 95.00 |
| Contextual feature | 93.81 | 95.88 | 95.91 | 95.88 |
| Hybrid feature | 94.09 | 96.22 | 96.24 | 96.21 |

### 5.3 Analysis of the Structural and Linguistic Contextual Feature

A comparison of the average evaluation metrics across the three spam detection model schemes demonstrates that the hybrid scheme is the most effective. This scheme achieves an accuracy of 93.22, precision of 95.25, recall of 95.12, and *F*-score of 95.12. The contextual feature scheme follows with accuracy, precision, recall, and *F*-score values of 92.56, 94.69, 94.48, and 94.49, respectively. The structural representation feature scheme shows the weakest performance among the three. The hybrid scheme's outstanding results, with precision, recall, and *F*-score values exceeding 95, can be attributed to its ability to leverage both structural and contextual features. While spam messages share similarities with conventional text classification, they also exhibit unique patterns, such as frequent use of uppercase and numeric characters, as shown in Tab. 2. However, understanding the semantic and linguistic aspects through contextual features proves more effective. Structural features alone are not always present in spam-indicated message texts, making them insufficient as the sole input. Overall, the hybrid scheme achieves the highest average evaluation metrics, as illustrated in Fig. 4. The hybrid scheme not only outperforms the other approaches in average metric values but also excels in multiple machine learning algorithms. For instance, logistic regression and random forest perform significantly better with the hybrid scheme compared to using only structural or contextual features. The hybrid scheme compensates for the limitations of linguistic representation features by integrating structural aspects, thereby providing a more comprehensive approach to spam detection.

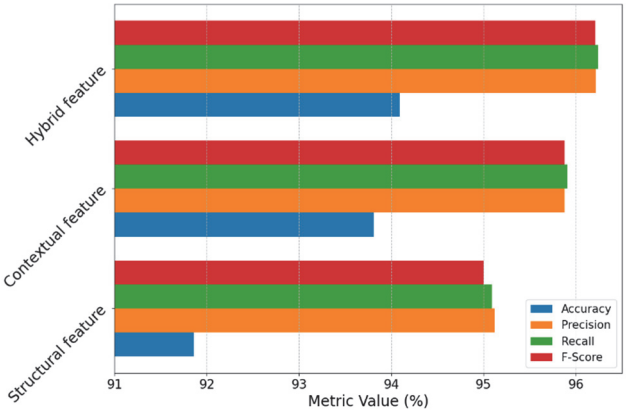


**Figure 4** Average of evaluation metrics on each scheme

### 5.4 Practical Implication and Limitations

The experimental results demonstrate the effectiveness of combining structural and contextual features for spam detection in SMS texts. This hybrid approach achieves robust performance by leveraging the embedded attributes of text data in messages to extract structural features while simultaneously capturing semantic and linguistic patterns through contextual features using a word embedding model. This dual advantage enables the hybrid model to maintain effectiveness even when structural features are absent, as contextual features can provide meaningful insights into spam message patterns. However, extracting contextual features requires additional computational resources and reliance on pre-trained word embedding models, whereas structural features can be obtained more easily.

This study focuses on text-based spam detection, comparing the performance of structural and contextual features. It also provides a comparative analysis of classifier algorithms to identify those that perform well in spam classification tasks. Despite these contributions, the study has certain limitations that should be acknowledged. First, this research is limited to text-based data and does not explore other approaches to spam detection, such as

those using sender-based attributes or behavioral features [34]. While the current study focuses on text data, integrating behavioral and sender-based features could further enhance detection in real-world scenarios. Second, the dataset used in this study consists exclusively of Indonesian SMS texts. This allows for a detailed exploration of the unique characteristics of the Indonesian language. Based on the number of slang word appearance, the average number of slang words per message is approximately 1.20, indicating that most messages contain at least one slang word. Nonetheless, this study provides meaningful insights into the comparative performance of structural and contextual features for spam detection. The findings highlight the practical utility of combining these features, especially in text-based spam detection scenarios, and lay the groundwork for future studies that could expand the approach to include additional data types or test its applicability in multilingual or cross-linguistic contexts.

## 6 CONCLUSIONS

In this paper, we present a comprehensive study of spam detection models that consider various features as input. These features include both contextual representations and structural attributes derived from message texts. Additionally, we propose and investigate the use of hybrid features, combining both structural and contextual representations, as input for spam detection models. Our experimental results demonstrate that relying solely on contextual representation features yields better performance for spam detection compared to structural features alone. However, combining structural and contextual features significantly outperforms single-feature schemes, achieving robust performance across multiple machine learning algorithms. This hybrid approach offers a practical solution for telecommunication providers by enhancing SMS filtering accuracy and reducing false positives. In addition, we found that the random forest model demonstrates robust performance in spam detection across various feature types. Moreover, the ability of the hybrid model to process unique linguistic patterns, such as slang commonly used in Indonesian texts, highlights its versatility and adaptability for multilingual deployments.

Finally, since the attributes of spam messages from around the world can be diverse, future studies should be conducted using different languages to determine the significant features of the spam detection model. Additionally, the next work should expand beyond text-based data to incorporate sender-based attributes and behavioral features, providing a more comprehensive approach to spam detection.

## 7 REFERENCES


[1] Community phone. (2022). *Spam Text Statistics for 2023*. https://www.communityphone.org/blogs/spam-text-statistics. Accessed 06-18-2024

[2] Xia, T. (2020). A constant time complexity spam detection algorithm for boosting throughput on rule-based filtering systems. *IEEE Access, 8*, 82653-82661. https://doi.org/10.1109/ACCESS.2020.2991328

[3] Tang, S., Mi, X., Li, Y., Wang, X., & Chen, K. (2022). Clues in tweets: twitter-guided discovery and analysis of SMS spam. *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, 2751-2764. https://doi.org/10.1145/3548606.3559351

[4] Rifat, M., Ahsan, M., Chowdhury, M., & Gomes, R. (2022). BERT against social engineering attack: phishing text detection. *Proceedings of* 2022 *IEEE International Conference on Electro Information Technology (eIT),* 1-6. https://doi.org/10.1109/eIT53891.2022.9813922

[5] Charanarur, P., Jain, H., Rao, G. S., Samanta, D., Sengar, S. S., & Hewage, C. T. (2023). Machine-learning-based spam mail detector. *SN Computer Science, 4*(858). https://doi.org/10.1007/s42979-023-02330-x

[6] Kuchipudi, B., Nannapaneni, R. T., & Liao, Q. (2020). Adversarial machine learning for spam filters. *Proceedings of* 15*th International Conference on Availability, Reliability and Security*, 1-6. https://doi.org/10.1145/3407023.3407079

[7] Batra, J., Jain, R., Tikkiwal, V. A., & Chakraborty, A. (2021). A comprehensive study of spam detection in e-mails using bio-inspired optimization techniques. *International Journal of Information Management Data Insights, 1*(1), 100006. https://doi.org/10.1016/j.jjimei.2020.100006

[8] Novo-Lourés, M., Ruano-Ordás, D., Pavón, R., Laza, R., Gómez-Meire, S., &Méndez, J. R. (2022). Enhancing representation in the context of multiple-channel spam filtering. *Information Processing & Management, 59*(2), 102812. https://doi.org/10.1016/j.ipm.2021.102812

[9] Oswald, C., Simon, S. E., & Bhattacharya, A. (2022). SpotSpam: Intention analysis-driven SMS spam detection using BERT embeddings. *ACM Transactions on the Web, 16*(3), 1-27. https://doi.org/10.1145/3538491

[10] Nasreen, G., Khan, M. M., Younus, M., Zafar, B., & Hanif, M. K. (2024). Email spam detection by deep learning models using novel feature selection technique and BERT. *Egyptian Informatics Journal, 26*, 100473. https://doi.org/10.1016/j.eij.2024.100473

[11] Duarte, J. M. & Berton, L. (2023). A review of semi-supervised learning for text classification. *Artificial Intelligence Review, 56*, 9401-9469. https://doi.org/10.1007/s10462-023-10393-8

[12] Osuagwu, E. C. (2020).An analysis of chat abbreviations and slangs of the students of the University of Port Harcourt. *English Linguistics Research, 9*(2), 22-31. https://doi.org/10.5430/elr.v9n2p22

[13] Abayomi-Alli, O., Misra, S., Abayomi-Alli, A., & Odusami, M. (2019). A review of soft techniques for SMS spam classification: Methods, approaches and applications. *Engineering Applications of Artificial Intelligence, 86*, 197-212. https://doi.org/10.1016/j.engappai.2019.08.024

[14] Iddrisu, W. A., Adjei-Gyabaa, S. K., & Akoto, I. (2023). Content-based spam classification of academic e-mails: a machine learning approach. *Proceedings of Advances in Information Communication Technology and Computing Proceedings of AICTC* 2022, 83-92. https://doi.org/10.1007/978-981-19-9888-1_7

[15] Ghosh, A. & Senthilrajan, A. (2023). Comparison of machine learning techniques for spam detection. *Multimedia Tools and Applications, 82*, 29227-29254. https://doi.org/10.1007/s11042-023-14689-3

[16] Faris, H., Al-Zoubi, A. M., Heidari, A. A., Aljarah, I., Mafarja, M., Hassonah, M. A., & Fujita, H. (2019). An intelligent system for spam detection and identification of the most relevant features based on evolutionary Random Weight Networks. *Information Fusion, 48*, 67-83. https://doi.org/10.1016/j.inffus.2018.08.002

[17] Majeed, H. A. & Bayati, M. A. (2018). Towards building an anti-spam for arabic SMS. *International Journal of Civil Engineering and Technology, 9*(10), 1402-1411.

[18] Rojas-Galeano, S. (2021). Using BERT encoding to tackle the mad-lib attack in SMS spam detection. *arXivpreprint*, arXiv:2107.06400v1.

[19] Raga, S. S. & Chaitra, B. L. (2022). A BERT model for SMS and Twitter spam ham classification and comparative study of machine learning and deep learning technique. *Proceedings of* 2022 *IEEE 7th International Conference on Recent Advances and Innovations in Engineering (ICRAIE)*, 355-359. https://doi.org/10.1109/ICRAIE56454.2022.10054285

[20] Andresini, G., Iovine, A., Gasbarro, R., Lomolino, M., Gemmis, Md., & Appice, A. (2022). Euphoria: A neural multi-view approach to combine content and behavioral features in review spam detection. *Journal of Computational Mathematics and Data Science, 3*, 100036. https://doi.org/10.1016/j.jcmds.2022.100036

[21] Uysal, A. K., Gunal, S., Ergin, S., & Gunal, E. S. (2013). The impact of feature extraction and selection on SMS spam filtering. *Elektronika ir Elektrotechnika, 19*(5), 67-72. https://doi.org/10.5755/j01.eee.19.5.1829

[22] Samsudin, N. M., Foozy, C. F. B. M., Alias, N., Shamala, P., Othman, N. F., & Din, W. I. S. W. (2019). Youtube spam detection framework using naïve bayes and logistic regression. *Indonesian Journal of Electrical Engineering and Computer Science, 14*(3), 1508-1517. https://doi.org/10.11591/ijeecs.v14.i3.pp1508-1517

[23] Niu, D., Wang, K., Sun, L., Wu, J., & Xu, X. (2020). Short-term photovoltaic power generation forecasting based on random forest feature selection and CEEMD: A case study. *Applied Soft Computing, 93*, 106389. https://doi.org/10.1016/j.asoc.2020.106389

[24] Fei, H., Fan, Z., Wang, C., Zhang, N., Wang, T., Chen, R., & Bai, T. (2022). Cotton Classification Method at the County Scale Based on Multi-Features and Random Forest Feature Selection Algorithm and Classifier. *Remote Sensing, 14*(4), 829. https://doi.org/10.3390/rs14040829

[25] Xiang, L., You, H., Guo, G., & Li, Q. (2023). Deep feature fusion for cold-start spam review detection. *J. Supercomputing, 79*(1), 419-434. https://doi.org/10.1007/s11227-022-04685-z

[26] 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.

[27] Chamlal, H., Kamel, H., & Ouaderhman, T. (2024). A hybrid multi-criteria meta-learner based classifier for imbalanced data. *Knowledge-Based Systems, 285*, 111367. https://doi.org/10.1016/j.knosys.2024.111367

[28] Truecaller. (2021).*Truecaller Insights: Top 20 Countries Affected by Spam Calls In 2021*. https://www.truecaller.com/blog/insights/top-20-countries-affected-by-spam-calls-in-2021. Accessed 27 May 2023

[29] Rahmi, F. & Wibisono, Y. (2016). *Aplikasi SMS Spam Filtering pada Android menggunakan Naive Bayes*. Thesis. Universitas Pendidikan Indonesia, Bandung, Indonesia.

[30] Tandra, V. C., Yowen, Y., Tanjaya, R., Santoso, W. L., & Qomariyah, N. N. (2021). Short message service filtering with natural language processing in indonesian language. *Proceedings of International Conference on ICT For Smart Society* 2021. https://doi.org/10.1109/ICISS53185.2021.9532503

[31] Reimers, N., Gurevych, I. (2019). Sentence-BERT: sentence embeddings using siamese BERT-networks. *arXiv preprint*, arXiv:1908.10084. https://doi.org/10.18653/v1/D19-1410

[32] Wilie, B., et al. (2020). IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding. *arXiv preprint*, arXiv:2009.05387.

[33] Zhang, X. & Liu, C. A. (2023). Model averaging prediction by K-fold cross-validation. *J. Econometrics, 235*(1), 280-301. https://doi.org/10.1016/j.jeconom.2022.04.007

[34] Saidat, M. R. A., Yerima, S. Y., & Shaalan, K. (2024). Advancements of SMS Spam Detection: A Comprehensive Survey of NLP and ML Techniques. *Procedia Computer Science, 244*, 248-259. https://doi.org/10.1016/j.procs.2024.10.198

**Contact information:**

**Ulin NUHA**, PhD
Dept. of Electronic Engineering,
National Kaohsiung University of Science and Technology,
Kaohsiung 80778, Taiwan
E-mail: i110152118@nkust.edu.tw

**Chih-Hsueh LIN**, Professor
(Corresponding author)
Dept. of Electronic Engineering,
National Kaohsiung University of Science and Technology,
Kaohsiung 80778, Taiwan
E-mail: cslin@nkust.edu.tw