Detection of Fraudulent and Promotional Messages in Bahasa Indonesia

Laela Citra Asih*
Informatics Engineering
Universitas Telkom, Bandung, West Java 40257
*laelacitrasih@gmail.com

Abstract—**The increasing misuse of messaging platforms in Indonesia for scams, fake promotions, and unwanted advertisements calls for reliable, language-specific detection systems. This paper introduces a lightweight, interpretable NLP-based solution that classifies Indonesian-language messages into three categories: normal, fraudulent, and promotional. Utilizing TF-IDF vectorization and Logistic Regression, our model demonstrates strong multi-class classification performance, with a macro F1-score of 0.93 and an accuracy of 94**

Fraud Detection, Text Classification, Logistic Regression, Spam Detection

I. Introduction

With the rapid growth of mobile phone usage in Indonesia, individuals are increasingly exposed to unsolicited messages, many of which contain fraudulent content or misleading promotions. Existing keyword-based filters often fall short, as malicious actors continually adapt the language used in scams. Consequently, there is a growing need for a context-aware system capable of understanding the linguistic structure and semantics of messages written in Bahasa Indonesia. This research aims to fill that gap by developing a machine learning model tailored to detect fraud-related and promotional text.

II. Proposed System

The proposed system is composed of three main components: data input, message classification, and result visualization. Users submit their message through a web-based interface built using Streamlit. The submitted text is preprocessed and transformed using the TF-IDF vectorizer. The Logistic Regression model then predicts the class of the message (normal, fraud, or promo). The system provides immediate feedback, indicating the classification result along with confidence scores and suggestions. Additional tabs offer insights into dataset composition, visualization of label distributions, and system overview.

A. Dataset

The research compiled a dataset of 1143 real and synthetic text messages written in Bahasa Indonesia. Each message was manually labeled based on its content: normal (legitimate communication), fraud (deceptive or scam messages), or promo (marketing/advertising).

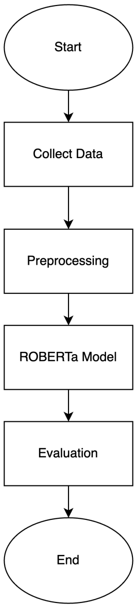

Fig. 1. System flowchart

B. Data Prepocessing

Before proceeding to model training, the data implemented a thorough data preprocessing routine to refine the message inputs. Each text message was first converted to lowercase to maintain consistency across all samples. Unnecessary characters—such as punctuation marks, numbers, and symbols—were stripped to prevent irrelevant tokens from influencing the model. We also removed extra white spaces and line breaks to simplify sentence structure. A crucial part of this process was the elimination of common stopwords in Bahasa Indonesia, such as “yang”, “dan”, or “di”, using a language-specific stopword list. These words often do not contribute meaningful information and can clutter the feature space. In certain preprocessing stages, messages were also tokenized to break down the sentences into individual words, improving the model’s ability to understand word-level patterns. This comprehensive cleaning strategy ensured that the data passed to the vectorizer was informative, structured, and representative of its intended meaning.

C. Logistic Regression

Logistic Regression was chosen as the classifier due to its simplicity, interpretability, and suitability for text classification tasks. It is a linear model that estimates the probability of a data point belonging to a certain class based on the weighted sum of its input features. In our case, the features are derived from TF-IDF scores, representing the importance of words and phrases within each message. The model uses a One-vs-Rest strategy for multi-class classification, which involves training one classifier per class while treating the remaining classes as a single group. This approach enables the model to make distinct predictions for each of the three message categories: normal, fraud, and promo.

To prevent overfitting and handle potential class imbalance, we employed L2 regularization and balanced class weighting. Additionally, we conducted basic hyperparameter tuning to optimize model performance. We experimented with different values of the regularization parameter C and evaluated the effects of using different solvers such as 'liblinear' and 'saga'. The best configuration was selected based on cross-validation results using stratified k-folds.

The final model was trained on 80% of the dataset, with stratified sampling ensuring that all classes were proportionally represented in both the training and test sets., with stratified sampling ensuring that all classes were proportionally represented in both the training and test sets. We selected Logistic Regression for its balance of speed and interpretability. The model was trained using a One-vs-Rest multi-class strategy. The dataset was split into training (80%) and testing (20%) subsets using stratified sampling to ensure equal representation across classes.

D. Evaluation

Model evaluation was conducted using both accuracy and F1 score as performance metrics. Accuracy measures the proportion of correctly predicted labels over the total number of predictions:

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} \times 100 \qquad (1)$$

F1 score is used to evaluate the balance between precision and recall, especially in imbalanced datasets. The F1 score is calculated as:

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \qquad (2)$$

Using the selected hyperparameters and validation set from the TweetEval benchmark, the model achieved an accuracy score of 69.4

III. Testing Result

To evaluate the robustness of the model, we conducted tests on unseen data samples. The model consistently returned accurate predictions across multiple examples. Table 1 presents several representative samples from each class to illustrate how the model performs in real-world scenarios.

The classifier achieved an overall accuracy of 94Normal: 0.98, Fraud: 0.91, Promo: 0.89. These results indicate that the model was able to correctly distinguish between normal communication and spam-like or fraudulent messages with a high degree of reliability. Messages labeled as ”Normal” were identified with near-perfect precision and recall, highlighting the model's ability to correctly recognize legitimate communication. Fraudulent messages also demonstrated strong performance, which is crucial given the potential consequences of failing to detect scams. Promotional messages, while slightly lower in performance, were still classified with acceptable accuracy, though some were mistakenly flagged as fraud and vice versa. This highlights the semantic similarities between persuasive marketing language and scam content.

The predictor demonstrated accurate classification in the majority of scenarios. However, contextually complex or nuanced tweets may yield unexpected outcomes, such as misclassifying clearly positive sentiments as neutral due to domain or language ambiguity. Further improvements may involve fine-tuning on domain-specific or multilingual datasets.

IV. Conclusion

This study demonstrates that traditional NLP techniques, when carefully adapted to the Indonesian language and messaging context, can effectively classify text messages into normal, fraudulent, and promotional categories. By applying TF-IDF vectorization and a Logistic Regression model, we achieved high performance with a macro F1-score of 0.93 and 94% accuracy—highlighting the model's potential for real-world deployment. The system's strength lies in its simplicity, efficiency, and interpretability, making it suitable for use in low-resource environments or integration into existing communication platforms.

Although the model performed well overall, some misclassifications occurred between fraud and promotional messages due to overlapping linguistic patterns. This points to a potential area for improvement through more advanced language models. Future work may involve scaling the dataset, exploring contextual models like IndoBERT, and evaluating the system's usability in real-time applications. Nevertheless, the current approach serves as a solid foundation for detecting and filtering unwanted or deceptive messages in Indonesian-language digital communication.