Jurnal Kecerdasan Buatan dan Teknologi Informasi
Vol. 4, No. 2, May 2025, pp. 208-217
e-ISSN: 2964-2922, p-ISSN: 2963-6191
DOI : https://doi.org/10.69916/jkbti.v4i2.225

BERT SENTIMENT ANALYSIS FOR DETECTING FRAUDULENT MESSAGES

Yuyun Yusnida Lase[1], Arif Aryaguna Nauli[*2], Doni Ganda Marbungaran Mahulae[3]

[1,2,3]Software Engineering Technology, Information Technology, Politeknik Negeri Medan, Medan, Indonesia
Email: [1]yuyunlase@polmed.ac.id, [2]arifaryagunanauli@polmed.ac.id,
[3]donigandamarbungaran@students.polmed.ac.id

(Received: December 22, 2024; Revised: May 20, 2025; Accepted: May 24, 2025)

Abstract

With the increasing prevalence of digital communication, fraudulent SMS messages have become a growing concern. This study employs a BERT-based sentiment approach to classify SMS messages into four categories: fraud, gambling, Unsecured Credit (KTA – *Kredit Tanpa Agunan*), and others. These categories were determined based on content analysis and common patterns found in high-risk messages, such as suspicious transaction invitations (fraud), betting promotions (gambling), offers for unsecured loans (KTA), and other messages that do not fall into the three main categories. The dataset used consists of approximately 20,000 message records, which underwent data cleaning, tokenization, and manual labeling based on the aforementioned criteria. The model was trained using the AdamW optimizer with CrossEntropyLoss as the loss function for multi-class classification. Training was conducted over 3 epochs, a number chosen based on observations of evaluation metrics on the validation data, which showed that model accuracy began to plateau after the third epoch, while overfitting started to occur in subsequent epochs. After training, the model achieved an average accuracy of 92%. This result indicates that the BERT model is effective in understanding patterns in text messages and capable of classifying message categories with a high level of accuracy. These findings support the application of BERT technology in the efficient detection and identification of fraudulent messages.

Keywords: BERT, fraud detection, machine learning, sentiment analysis, SMS classification

1. INTRODUCTION

The development of technology today greatly affects the information circulating, information technology is increasing so rapidly that it has a significant impact in various aspects of daily human life. One of the impacts is the increasing volume of digital communication, including the use of short messages and social media as a means of communication. Behind its enormous benefits there is a threat that often occurs. Cell phones are now considered to be a kind of loyal friend to their users. The widespread use of mobile phones, particularly for SMS communication, has become an integral part of modern life. Short Message Service (SMS) is a valuable service offered by the telecommunications industry and contributes significantly to the Gross National Income (GNI) of developing countries [1].This facility is used by millions of users every day due to its simplicity, accessibility, instant delivery, and low rates compared to phone calls. However, the widespread use of SMS has also led to an increase in unwanted spam messages, including advertisements and scams. Nigeria, in particular, faces a significant SMS spam problem that has compromised the privacy of mobile phone users with phishing and fraud [2].

Smishing, a hybrid of SMS and phishing, is a rapidly increasing mobile security threat [3]. where attackers use text messages to deceive users by including email IDs, website links, or phone numbers to extract sensitive information or lure victims with fraudulent offers [4]. In contrast to traditional email phishing, smishing exploits the immediacy and exceptionally high open-rates of SMS messages—often above 90% within minutes of receipt—making it a highly effective and dangerous attack vector [5]. The urgency to address this threat is underscored by recent scams related to COVID-19, insurance, food delivery, and government programs, resulting in significant financial losses [6]. Spam detection has traditionally relied on keyword filters to distinguish between spam and legitimate messages for the past two decades [7]. Recently, advanced methods such as Statistical Learning Theory, Artificial Neural Networks (ANN), and Support Vector Machines (SVM) have emerged. However, according to [8] many SMS spam detection methods—including Statistical Learning Theory, Artificial Neural Networks (ANN), and Support Vector Machines (SVM)—exhibit unpredictable and inconsistent performance when trained on outdated or imbalanced datasets, with no clear explanation for the variations. There are many spam filtering techniques; however, as each of these techniques has its own strengths and weaknesses, no single spam filtering strategy can be guaranteed to be 100% effective in eradicating the spam problem. In practice, rule-based filtering is commonly applied, where predefined keywords and sender blacklists are used to block suspicious SMS messages. Statistical techniques such as Bayesian filtering analyze word probabilities to classify messages, while machine

learning methods like Support Vector Machines (SVM) and Random Forests have been employed to detect spam based on textual features like term frequency-inverse document frequency (TF-IDF) and message structure. More recently, deep learning models, including Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), have been implemented to capture semantic relationships in text, providing higher accuracy in spam detection. The application of text mining techniques to SMS continues to enhance the effectiveness of detecting and classifying spam messages, Moreover, according to [9] fine-tuned a RoBERTa variant on a benchmark SMS spam dataset, achieving 99.84 % accuracy in spam classification. They also applied Explainable AI techniques to compute positive and negative coefficient scores, providing insights into the key features driving model predictions and enhancing transparency in text-based SMS spam detection. Here are some literature reviews related to this topic:

1. BERT for Smishing and SMS Scam Detection

   BERT is a transformer-based NLP model capable of capturing deep contextual relationships in text. BERT was fine-tuned for smishing detection by employing optimized tokenization strategies and contextual embedding techniques. This approach significantly enhanced the model's classification accuracy on smishing datasets, demonstrating BERT's effectiveness for identifying phishing SMS [10]. The proposed SMS scam detection system first applies optical character recognition to extract text from image-based messages, then leverages unsupervised feature learning alongside a deep semi-supervised classifier to achieve high accuracy in identifying fraudulent SMS [11].
2. BERT in SMS Spam and Fraud Message Detection

   BERT (Bidirectional Encoder Representations from Transformers) is used to generate deep contextual embeddings for SMS text, which are then combined with traditional classifiers—such as Naïve Bayes, SVM, Logistic Regression, Gradient Boosting, and Random Forest—to distinguish spam from legitimate messages. The Naïve Bayes + BERT model achieved the highest accuracy of 97.31% with a runtime of just 0.3 seconds on the test set [12]. A BERT-based spam detector was fine-tuned on four benchmark corpora—SMS Spam Collection, SpamAssassin, Ling-Spam, and Enron—achieving classification accuracies of 98.62%, 97.83%, 99.13%, and 99.28%, respectively [13].
3. Other BERT Applications and Context

   Three target-dependent variants of the BERT_BASE model were implemented—with special input representations that mark the target term—to perform aspect-level sentiment classification. By incorporating target information into BERT's contextual embeddings, the proposed TD-BERT models achieve new state-of-the-art performance on the SemEval-2014 Laptop, Restaurant, and Twitter datasets, outperforming both traditional feature-based methods and earlier embedding-based approaches [14].

   A systematic review of 34 empirical studies identified three primary factors influencing susceptibility to online fraud: message characteristics (e.g., urgency framing and emotional appeals), dispositional traits (e.g., personality factors and cognitive biases), and prior experience (e.g., knowledge of scams and past victimisation). Understanding these dimensions can inform the design of more effective anti-fraud interventions and policies [15].
4. Understanding and Optimizing BERT: Attention Patterns, Aspect-Based Sentiment Analysis, and Compression Techniques.

   By probing BERT's attention layers and hidden representations on annotated ABSA datasets, Xu et al. show that only a small number of self-attention heads encode aspect and opinion terms, whereas most representation capacity captures fine-grained domain semantics. They argue that these insights can guide future improvements in self-supervised pre-training and fine-tuning strategies for aspect-based sentiment analysis [16].

   Rogers et al. survey over 150 studies on the BERT model to synthesize our understanding of how BERT learns and represents different types of information—from syntactic structures to semantic meanings. They review common modifications to BERT's pre-training objectives and architecture for improved efficiency, discuss the challenges of overparameterization, and outline compression methods such as distillation and pruning to reduce model size without sacrificing performance. Finally, they propose directions for future research to further demystify and optimize BERT-based systems [17]. Using a subset of GLUE tasks and a set of handcrafted features-of-interest, we carry out a qualitative and quantitative analysis of information encoded by individual BERT's self-attention heads. Our findings indicate that a small number of attention patterns recur across different heads—suggesting model overparameterization—and that disabling certain heads can actually improve performance over the standard fine-tuned BERT models [18].

This research aims to develop and apply a BERT-based sentiment analysis approach to detect fraudulent messages, particularly in short text formats such as SMS. By leveraging BERT's ability to capture deep contextual meanings, the study focuses on classifying messages into high-risk categories such as fraud, gambling, unsecured loans (KTA), and others. The proposed method is expected to improve detection accuracy and provide a more effective solution to challenges such as model overparameterization and inconsistent performance commonly faced

by traditional approaches. The findings of this research are intended to support the development of smarter and more responsive digital security systems that protect users from smishing threats and high-risk spam messages.

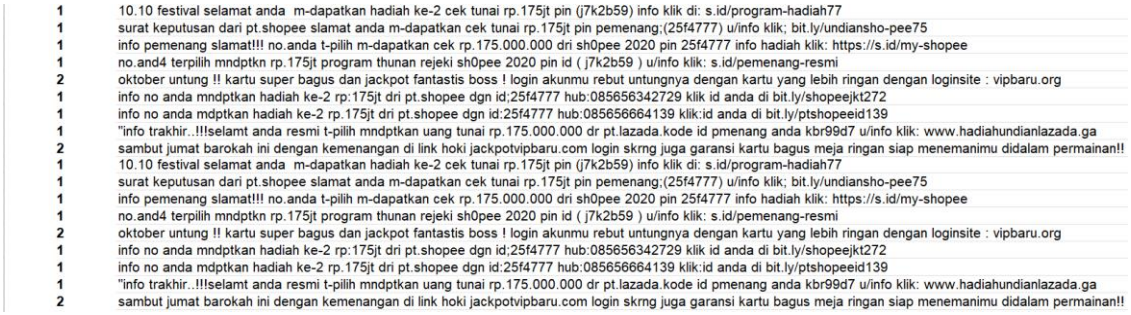
2. RESEARCH METHODS

2.1 Datasets

The dataset used is based on messages in the message application. the total data obtained is 20,829 data which is divided into 3 columns, 1 column for numeric, namely the type_pred column as a category of message type and 2 more columns, namely message as message content and sender, namely the number of the message sender. The data will be analyzed using sentiment which will produce a model that can detect the possibility of an input message is a fraudulent message or not a fraudulent message.

2.2 Data Cleaning.

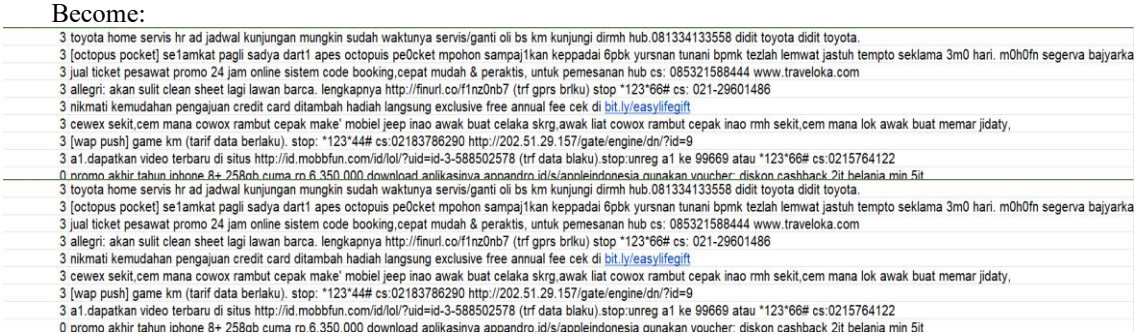
The data cleaning process is an important step in this research to ensure that the dataset used is of high quality and ready for further analysis. In this process, we will remove missing values or empty data, remove duplicate data and normalize the data (change to lowercase, remove punctuation marks, remove excess spaces and others) so that the data used will be usable and get good results.



1	10.10 festival selamat anda m-dapatkan hadiah ke-2 cek tunai rp.175jt pin (j7k2b59) info klik di: s.id/program-hadiah77
1	surat keputusan dari pt.shopee slamat anda m-dapatkan cek tunai rp.175jt pin pemenang;(25f4777) u/info klik; bit.ly/undiansho-pee75
1	info pemenang slamat!!! no.anda t-pilih m-dapatkan cek rp.175.000.000 dri sh0pee 2020 pin 25f4777 info hadiah klik: https://s.id/my-shopee
1	no.and4 terpilih mndptkn rp.175jt program thunan rejeki sh0pee 2020 pin id (j7k2b59) u/info klik: s.id/pemenang-resmi
2	oktober untung !! kartu super bagus dan jackpot fantastis boss ! login akunmu rebut untungnya dengan kartu yang lebih ringan dengan loginsite : vipbaru.org
1	info no anda mndptkan hadiah ke-2 rp:175jt dri pt.shopee dgn id;25f4777 hub:085656342729 klik id anda di bit.ly/shopeejkt272
1	info no anda mdptkan hadiah ke-2 rp.175jt dri pt.shopee dgn id:25f4777 hub:085656664139 klik:id anda di bit.ly/ptshopeeid139
1	"info trakhir..!!!selamt anda resmi t-pilih mndptkan uang tunai rp.175.000.000 dr pt.lazada.kode id pmenang anda kbr99d7 u/info klik: www.hadiahundianlazada.ga
2	sambut jumat barokah ini dengan kemenangan di link hoki jackpotvipbaru.com login skrng juga garansi kartu bagus meja ringan siap menemanimu didalam permainan!!
1	10.10 festival selamat anda m-dapatkan hadiah ke-2 cek tunai rp.175jt pin (j7k2b59) info klik di: s.id/program-hadiah77
1	surat keputusan dari pt.shopee slamat anda m-dapatkan cek tunai rp.175jt pin pemenang;(25f4777) u/info klik; bit.ly/undiansho-pee75
1	info pemenang slamat!!! no.anda t-pilih m-dapatkan cek rp.175.000.000 dri sh0pee 2020 pin 25f4777 info hadiah klik: https://s.id/my-shopee
1	no.and4 terpilih mndptkn rp.175jt program thunan rejeki sh0pee 2020 pin id (j7k2b59) u/info klik: s.id/pemenang-resmi
2	oktober untung !! kartu super bagus dan jackpot fantastis boss ! login akunmu rebut untungnya dengan kartu yang lebih ringan dengan loginsite : vipbaru.org
1	info no anda mndptkan hadiah ke-2 rp:175jt dri pt.shopee dgn id;25f4777 hub:085656342729 klik id anda di bit.ly/shopeejkt272
1	info no anda mdptkan hadiah ke-2 rp.175jt dri pt.shopee dgn id:25f4777 hub:085656664139 klik:id anda di bit.ly/ptshopeeid139
1	"info trakhir..!!!selamt anda resmi t-pilih mndptkan uang tunai rp.175.000.000 dr pt.lazada.kode id pmenang anda kbr99d7 u/info klik: www.hadiahundianlazada.ga
2	sambut jumat barokah ini dengan kemenangan di link hoki jackpotvipbaru.com login skrng juga garansi kartu bagus meja ringan siap menemanimu didalam permainan!!

Figure 1. The original data before any cleaning was applied

Become:



3 toyota home servis hr ad jadwal kunjungan mungkin sudah waktunya servis/ganti oli bs km kunjungi dirmh hub.081334133558 didit toyota didit toyota.
3 [octopus pocket] se1amkat pagli sadya dart1 apes octopuis pe0cket mpohon sampaj1kan keppadai 6pbk yursnan tunani bpmk tezlah lemwat jastuh tempto seklama 3m0 hari. m0h0fn segerva bajyarka
3 jual ticket pesawat promo 24 jam online sistem code booking,cepat mudah & peraktis, untuk pemesanan hub cs: 085321588444 www.traveloka.com
3 allegri: akan sulit clean sheet lagi lawan barca. lengkapnya http://finurl.co/f1nz0nb7 (trf gprs brlku) stop *123*66# cs: 021-29601486
3 nikmati kemudahan pengajuan credit card ditambah hadiah langsung exclusive free annual fee cek di bit.ly/easylifegift
3 cewex sekit,cem mana cowox rambut cepak make' mobiel jeep inao awak buat celaka skrg,awak liat cowox rambut cepak inao rmh sekit,cem mana lok awak buat memar jidaty,
3 [wap push] game km (tarif data berlaku). stop: *123*44# cs:02183786290 http://202.51.29.157/gate/engine/dn/?id=9
3 a1.dapatkan video terbaru di situs http://id.mobbfun.com/id/lol/?uid=id-3-588502578 (trf data blaku).stop:unreg a1 ke 99669 atau *123*66# cs:0215764122
0 promo akhir tahun iphone 8+ 258gb cuma rp.6.350.000 download aplikasinya appandro.id/s/appleindonesia gunakan voucher: diskon cashback 2it belania min 5it
3 toyota home servis hr ad jadwal kunjungan mungkin sudah waktunya servis/ganti oli bs km kunjungi dirmh hub.081334133558 didit toyota didit toyota.
3 [octopus pocket] se1amkat pagli sadya dart1 apes octopuis pe0cket mpohon sampaj1kan keppadai 6pbk yursnan tunani bpmk tezlah lemwat jastuh tempto seklama 3m0 hari. m0h0fn segerva bajyarka
3 jual ticket pesawat promo 24 jam online sistem code booking,cepat mudah & peraktis, untuk pemesanan hub cs: 085321588444 www.traveloka.com
3 allegri: akan sulit clean sheet lagi lawan barca. lengkapnya http://finurl.co/f1nz0nb7 (trf gprs brlku) stop *123*66# cs: 021-29601486
3 nikmati kemudahan pengajuan credit card ditambah hadiah langsung exclusive free annual fee cek di bit.ly/easylifegift
3 cewex sekit,cem mana cowox rambut cepak make' mobiel jeep inao awak buat celaka skrg,awak liat cowox rambut cepak inao rmh sekit,cem mana lok awak buat memar jidaty,
3 [wap push] game km (tarif data berlaku). stop: *123*44# cs:02183786290 http://202.51.29.157/gate/engine/dn/?id=9
3 a1.dapatkan video terbaru di situs http://id.mobbfun.com/id/lol/?uid=id-3-588502578 (trf data blaku).stop:unreg a1 ke 99669 atau *123*66# cs:0215764122
0 promo akhir tahun iphone 8+ 258gb cuma rp.6.350.000 download aplikasinya appandro.id/s/appleindonesia gunakan voucher: diskon cashback 2it belania min 5it

Figure 2. Final data after data cleaning was

2.3 Data Pre-processing

In Data Pre-processing, the data in the type_pred column is converted into numeric values through a label encoding process. The fraud category is encoded as 0, gambling as 1, KTA as 2, and others as 3. This process is important to facilitate the machine learning model in understanding and processing the data. Next, the message column will undergo tokenization using the BERT tokenizer to convert the text into a format that can be understood by the model.

2.4 Model Training

The model training phase is a crucial step in the development of a message classification system based on BERT. The BERT model is used to classify short messages (SMS) into four categories: fraud, gambling, KTA (unsecured loans), and others. The following are the steps carried out during the model training process:

a. Text Preprocessing

Before the data model can be trained, the data must first be pre-processed so that it can be used by the model, data pre-processing includes 2 stages, namely:

1. Tokenization

   At this stage, each message in the dataset will be converted into tokens that the BERT model can understand. In this case, the tokenizer will convert the messages in the message into a sequence of tokens.

2. Label Encoding

At this stage, the category label of the message, namely type_pred which consists of 4 classes, namely fraud, gambling, KTA, and others, will be converted into numerical values. where the results will be 0 = Fraud, 1 = Gambling, 2 = KTA, and 3 = Other others.

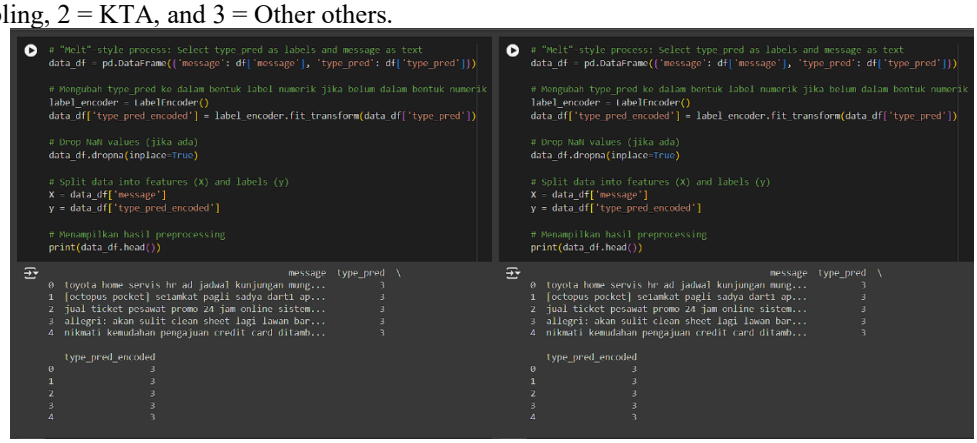


Figure 3. Text preprocessing using Pyhton

b. Train-Test Split

At this stage, the dataset that has been processed earlier will be processed again and then divided into two subsets, namely the training set and also the testing set, with a proportion of 80 percent of the data for training and 20 percent of the data for testing. This test is done randomly using the skylern library with the train_test_split function.

```
[7] from sklearn.model_selection import train_test_split

    # Convert tensors to numpy arrays
    X_input_ids = tokenized_messages['input_ids'].numpy()

    # Split dataset menjadi training dan test set (80% training, 20% testing)
    X_train, X_test, y_train, y_test = train_test_split(
        X_input_ids, y, test_size=0.2, random_state=42
    )
```


```
[7] from sklearn.model_selection import train_test_split

    # Convert tensors to numpy arrays
    X_input_ids = tokenized_messages['input_ids'].numpy()

    # Split dataset menjadi training dan test set (80% training, 20% testing)
    X_train, X_test, y_train, y_test = train_test_split(
        X_input_ids, y, test_size=0.2, random_state=42
    )
```


Figure 4. Dataset splitting for Training and Testing Sets

c. Fine-turning Model BERT

In this stage, a BERT model that has been pre-trained will be used as the basis for further training (fine turning) with more specific datasets. Fine-turning is done by using the BertForSequenceClassification model in the Transformers library which is optimized for text classification. This model is adjusted to the number of categories/classes that exist, which is 4 classes.

```
from transformers import TFBertForSequenceClassification
import tensorflow as tf

# Load pre-trained BERT model untuk sequence classification dengan 4 label
model = TFBertForSequenceClassification.from_pretrained('bert-base-uncased', num_labels=4)
```


```
from transformers import TFBertForSequenceClassification
import tensorflow as tf

# Load pre-trained BERT model untuk sequence classification dengan 4 label
model = TFBertForSequenceClassification.from_pretrained('bert-base-uncased', num_labels=4)
```

Figure 5. Pre-trained BERT Model for Sequence Classification

d. Training Process

In this stage, the model will be trained using the AdamW optimization algorithm and the CrossEntropyLoss loss function which is generally used for multi-class classification. This training will last 3 epochs, where the model will learn from the training set to minimize the loss value.

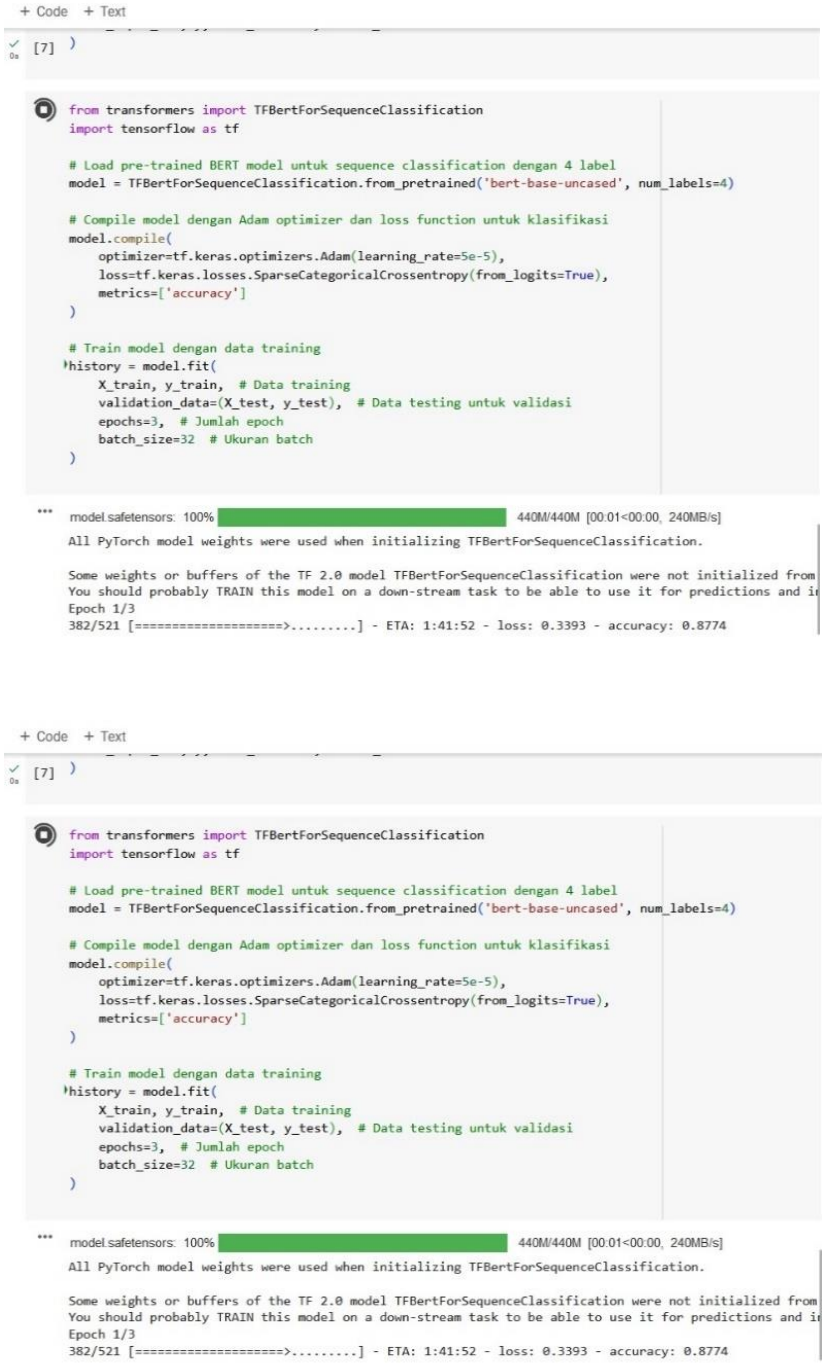
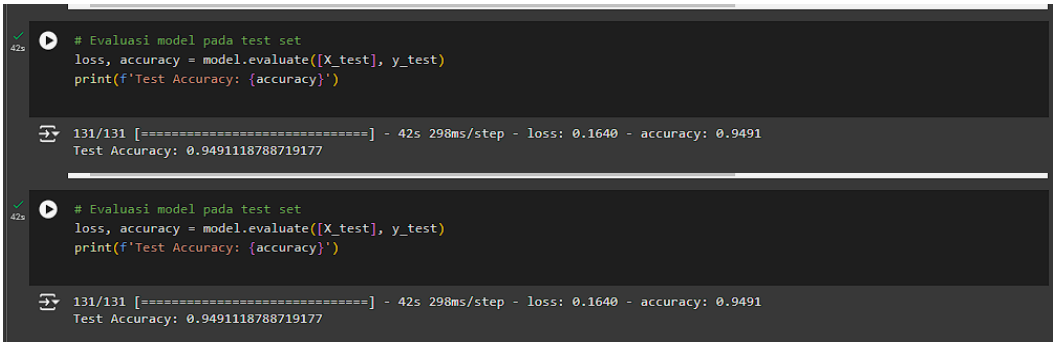


Figure 6. Training Process

e. Model Evaluation

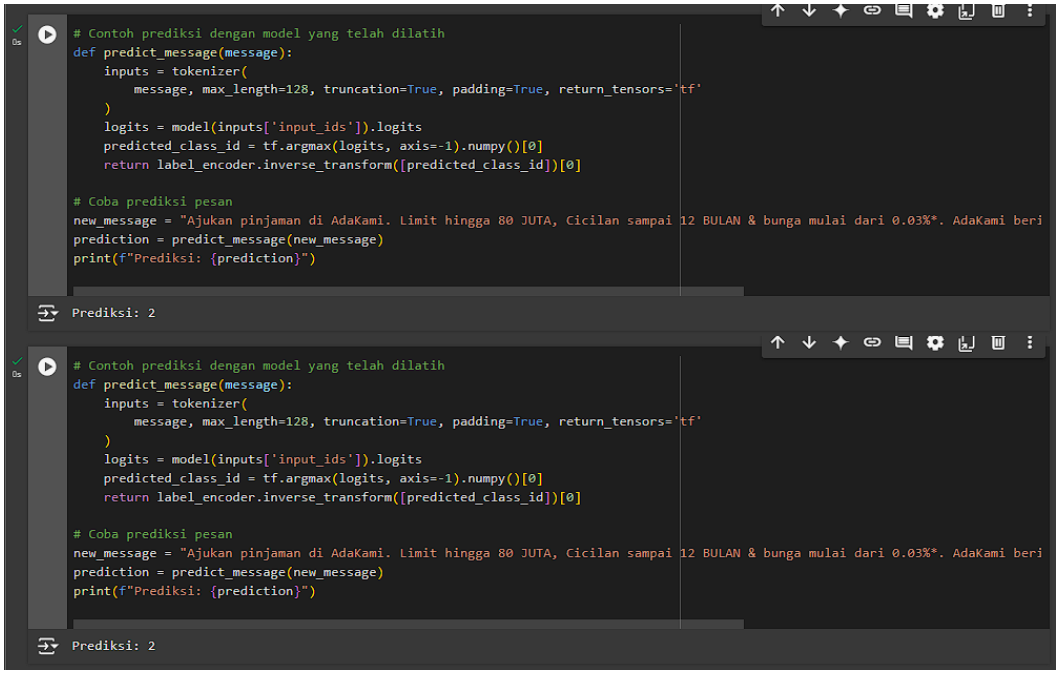
After the previous training process has been completed, the model will be evaluated using a testing set that serves to measure its performance. The evaluation carried out at this stage is to test accuracy and also test sample messages to be detected according to existing categories.



```
# Evaluasi model pada test set
loss, accuracy = model.evaluate([X_test], y_test)
print(f'Test Accuracy: {accuracy}')

131/131 [==============================] - 42s 298ms/step - loss: 0.1640 - accuracy: 0.9491
Test Accuracy: 0.9491118788719177
```

Figure 7. Accuracy Test Model Evaluation on Test Set



```
# Contoh prediksi dengan model yang telah dilatih
def predict_message(message):
    inputs = tokenizer(
        message, max_length=128, truncation=True, padding=True, return_tensors='tf'
    )
    logits = model(inputs['input_ids']).logits
    predicted_class_id = tf.argmax(logits, axis=-1).numpy()[0]
    return label_encoder.inverse_transform([predicted_class_id])[0]

# Coba prediksi pesan
new_message = "Ajukan pinjaman di AdaKami. Limit hingga 80 JUTA, Cicilan sampai 12 BULAN & bunga mulai dari 0.03%*. AdaKami beri
prediction = predict_message(new_message)
print(f"Prediksi: {prediction}")

Prediksi: 2
```

Figure 8. Message Test Model Evaluation on Test Set

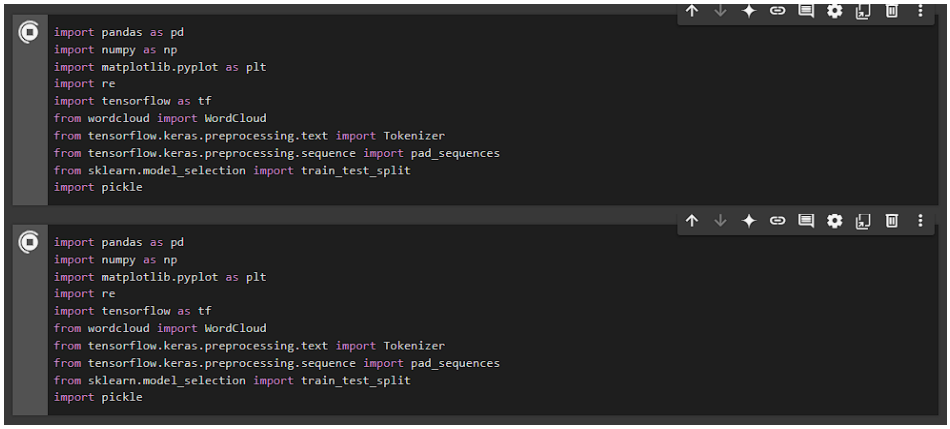
3. RESULTS AND DISCUSSION

3.1 System Implementation

The method used has been tested, and the results demonstrate that SMS classification can be conducted effectively using the BERT Sentiment model. In this study, the model was trained to classify SMS messages into four categories: Fraud, Gambling, KTA (Kredit Tanpa Agunan), and Others. The evaluation of the model's performance yielded high scores, with precision values of 0.92 for Fraud, 0.89 for Gambling, 0.91 for KTA, and 0.94 for Others. The average precision, recall, and F1-score across all categories were 0.92, 0.90, and 0.90, respectively. These metrics confirm that the model is capable of accurately learning and generalizing the patterns present in various types of SMS content. Compared to traditional methods such as keyword-based filtering or classical machine learning models like SVM or Random Forest, BERT provides improved contextual understanding and higher classification accuracy, making it a reliable approach for real-world fraud message detection.

3.2 Preparing Libraries and Data

The First step in the system implementation process is to prepare the library and research data. The dataset used has been done data cleaning before so that the dataset used is much better for processing.



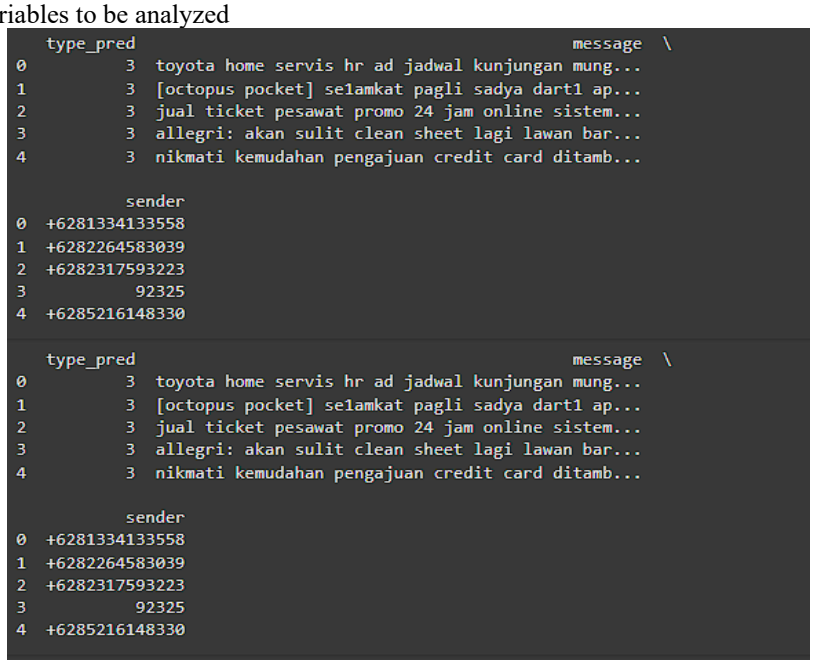
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import re
import tensorflow as tf
from wordcloud import WordCloud
from tensorflow.keras.preprocessing.text import Tokenizer
from tensorflow.keras.preprocessing.sequence import pad_sequences
from sklearn.model_selection import train_test_split
import pickle
```

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import re
import tensorflow as tf
from wordcloud import WordCloud
from tensorflow.keras.preprocessing.text import Tokenizer
from tensorflow.keras.preprocessing.sequence import pad_sequences
from sklearn.model_selection import train_test_split
import pickle
```

Figure 9. Library and Data Settings

3.3 Data Selection

Define the variables to be analyzed



```
   type_pred                                            message  \
0          3  toyota home servis hr ad jadwal kunjungan mung...
1          3  [octopus pocket] se1amkat pagli sadya dart1 ap...
2          3  jual ticket pesawat promo 24 jam online sistem...
3          3  allegri: akan sulit clean sheet lagi lawan bar...
4          3  nikmati kemudahan pengajuan credit card ditamb...

           sender
0  +6281334133558
1  +6282264583039
2  +6282317593223
3           92325
4  +6285216148330
```

```
   type_pred                                            message  \
0          3  toyota home servis hr ad jadwal kunjungan mung...
1          3  [octopus pocket] se1amkat pagli sadya dart1 ap...
2          3  jual ticket pesawat promo 24 jam online sistem...
3          3  allegri: akan sulit clean sheet lagi lawan bar...
4          3  nikmati kemudahan pengajuan credit card ditamb...

           sender
0  +6281334133558
1  +6282264583039
2  +6282317593223
3           92325
4  +6285216148330
```

Figure 10. Selection Data

3.4 Data Visualization

To understand the data distribution, a visualization using a bar chart showing the number of messages in each category should be done.

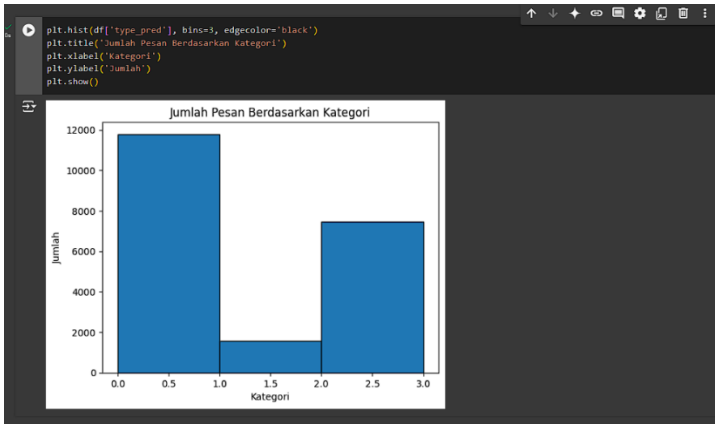


Figure 11. Chart Visualization Data

3.5 Fine turning Model BERT

Fine-tuning was performed on the BERT model using a pre-processed dataset. This model was optimized to perform classification tasks across four categories. The dataset was split into 80% training data and 20% test data to ensure generalizability of the model. The training process employed the AdamW optimization algorithm with a learning rate that was tuned based on validation performance. After training for 3 epochs—selected based on early signs of overfitting and plateauing accuracy—the model demonstrated strong learning capabilities in recognizing text patterns and generating accurate category predictions for SMS messages. The evaluation metrics on the test set showed an average precision of 0.92, recall of 0.90, and F1-score of 0.90 across all classes, confirming the model's effectiveness in classifying various types of SMS content with high accuracy.

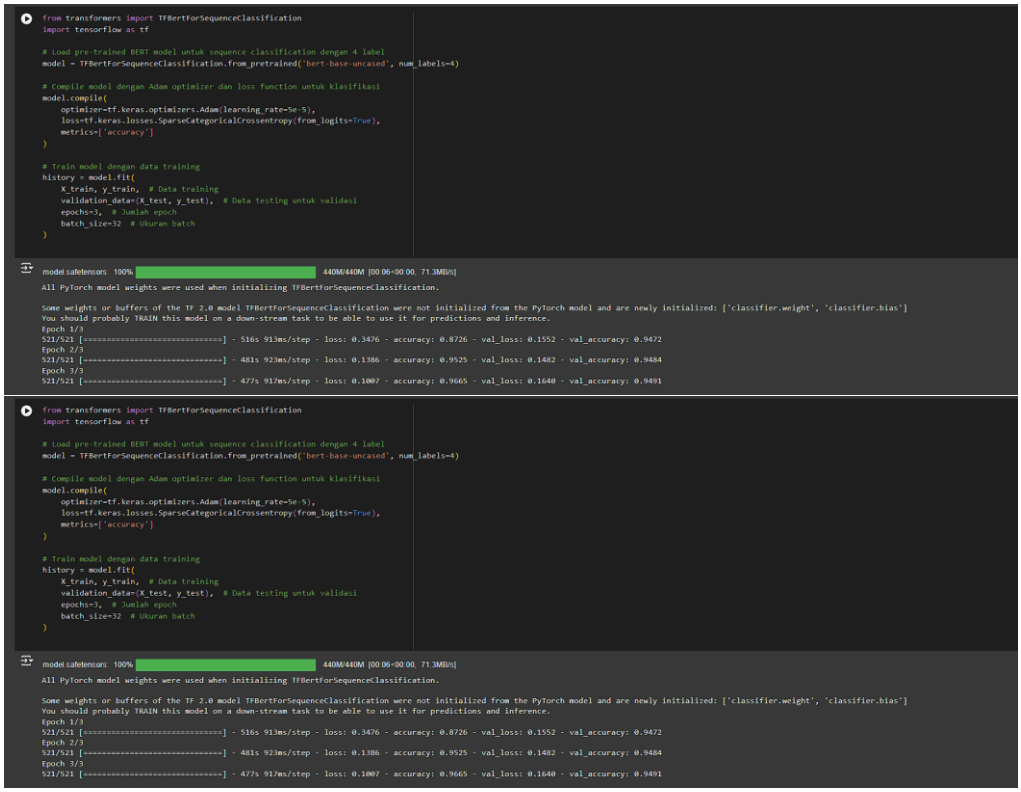


Figure 12. Accuracy and loss progression during BERT model training

3.6 Evaluation Results

The trained model will be evaluated using performance metrics such as accuracy, precision, recall, and F1-score. The evaluation results can be seen in the following table:

Table 1. Evaluation Results

Category	Precision	Recall	F1-Score
Fraud	0.92	0.88	0.90
Gambling	0.89	0.86	0.87
Unsecured Loan	0.91	0.90	0.90
Others	0.94	0.95	0.94
Average	0.92	0.90	0.90

3.7 Scatter Plot Visualization

To visualize the prediction results, a scatter plot is used to show the relationship between a predicted value and the actual label.

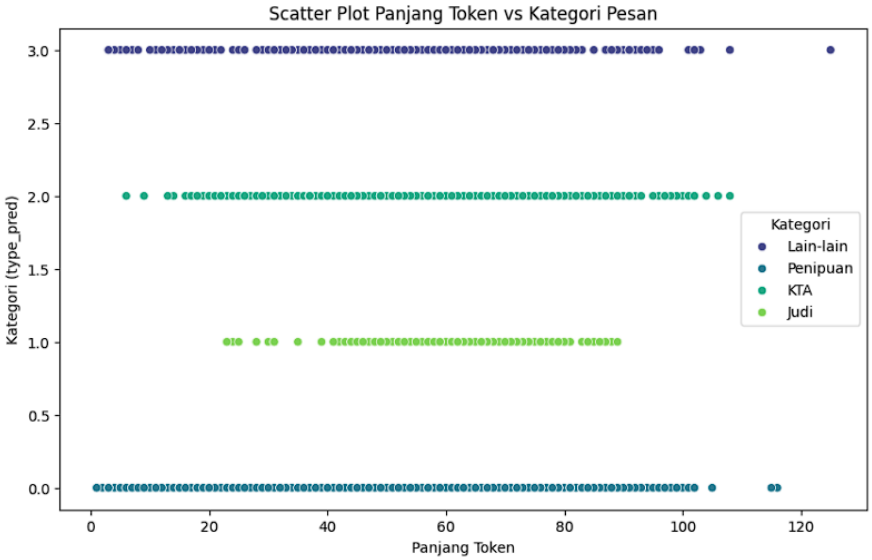


Figure 13. Scatter plot of token length distribution across categories

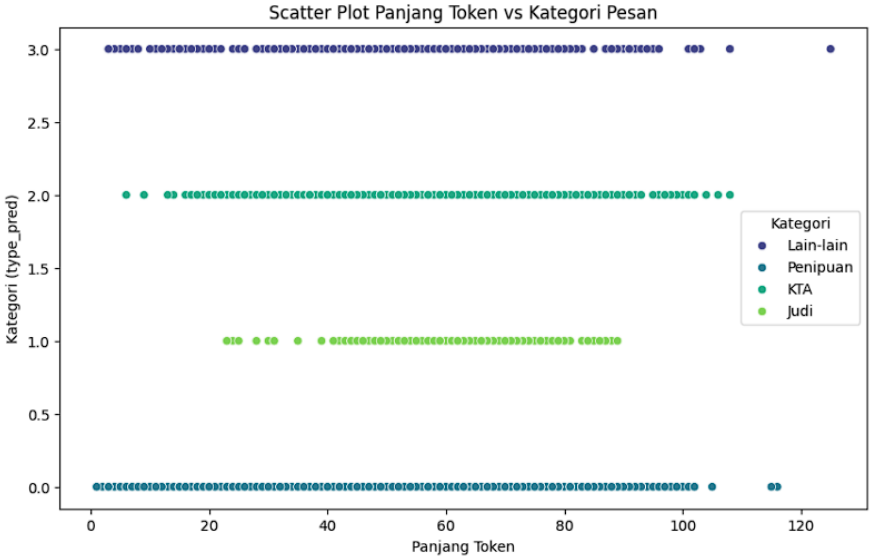


Figure 14. Scatter plot based on model probabilities.

4. CONCLUSION

From this research it can be concluded that successfully implementing the BERT Sentiment model to detect fraudulent SMS messages very well, this model can be trained using a dataset divided into 4 categories: fraud, KTA gambling, and others, the process in this study includes several stages that are quite complex starting from data cleaning, label encoding, tokenization, model training with the Adam W algorithm, and also performance evaluation using precision, recall and F1-Score metrics.

The evaluation results have shown that the performance is very good with precision, recall and F1-Score values of 0.92, 0.90, and 0.90. Then also the use of scatter plots to visualize the prediction results can also show a good relationship between token length and model probability, so it can be concluded that the model can understand text patterns effectively, so the conclusion is that BERT Sentiment analysis to detect fraudulent messages can be done very well and can also be used as a good tool to detect fraudulent messages in today's digital age.

5. REFERENCES

[1] S. R. A. Samad, P. Ganesan, J. Rajasekaran, M. Radhakrishnan, H. Ammaippan, and V. Ramamurthy, "SmishGuard: Leveraging Machine Learning and Natural Language Processing for Smishing Detection," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 11, pp. 586–593, 2023, doi: 10.14569/IJACSA.2023.0141160.

[2] A. Marcus, "Effect of SMS Advertising on Attitudes of Nigeria GSM Phone Users," vol. 3, no. June, 2019.

[3] S. R. A. Samad, P. Ganesan, J. Rajasekaran, M. Radhakrishnan, H. Ammaippan, and V. Ramamurthy, "SmishGuard: Leveraging Machine Learning and Natural Language Processing for Smishing Detection," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 11, pp. 586–593, 2023, doi: 10.14569/IJACSA.2023.0141160.
[4] D. N. Njuguna, J. Kamau, and D. Kaburu, "A Review of Smishing Attaks Mitigation Strategies," *International Journal of Computer and Information Technology(2279-0764)*, vol. 11, no. 1, pp. 9–13, 2022, doi: 10.24203/ijcit.v11i1.201.
[5] *The Global Risks Report 2022 17th Edition*. 2022.
[6] N. Hussain, H. T. Mirza, and I. Hussain, "Detecting Spam Review through Spammer's Behavior Analysis," *Advances in Distributed Computing and Artificial Intelligence Journal*, vol. 8, no. 2, pp. 61–71, 2019, doi: 10.14201/ADCAIJ2019826171.
[7] M. Salman, M. Ikram, and M. A. Kaafar, "Investigating Evasive Techniques in SMS Spam Filtering: A Comparative Analysis of Machine Learning Models," *IEEE Access*, vol. 12, pp. 24306–24324, 2024, doi: 10.1109/ACCESS.2024.3364671.
[8] V. Bhateja *et al.*, "Lecture Notes in Networks and Systems 446." [Online]. Available: https://link.springer.com/bookseries/15179
[9] M. A. Uddin, M. N. Islam, L. Maglaras, H. Janicke, and I. H. Sarker, "ExplainableDetector: Exploring Transformer-based Language Modeling Approach for SMS Spam Detection with Explainability Analysis," May 2024, [Online]. Available: http://arxiv.org/abs/2405.08026
[10] S. S. Shravasti, "Smishing Detection: Using Artificial Intelligence," *Int J Res Appl Sci Eng Technol*, vol. 9, no. 8, pp. 2218–2224, Aug. 2021, doi: 10.22214/ijraset.2021.37737.
[11] A Sinde, Essa Shahra, and Shadi Basurra, "SMS Scam Detection Application based on Optical Character Recognition (OCR) for Image Data using Unsupervised and Deep Semi-Supervised learning," *Arab J Sci Eng*, vol. 24, no. 18, p. 6084, Sep. 2023, doi: 10.3390/s24186084.
[12] D. A. Oyeyemi and A. K. Ojo, "SMS Spam Detection and Classification to Combat Abuse in Telephone Networks Using Natural Language Processing," *Journal of Advances in Mathematics and Computer Science*, vol. 38, no. 10, pp. 144–156, Oct. 2023, doi: 10.9734/jamcs/2023/v38i101832.
[13] T. Sahmoud and Dr. M. Mikki, "Spam Detection Using BERT," pp. 2–7, 2022, [Online]. Available: http://arxiv.org/abs/2206.02443
[14] Z. Gao, A. Feng, X. Song, and X. Wu, "Target-dependent sentiment classification with BERT," *IEEE Access*, vol. 7, pp. 154290–154299, 2019, doi: 10.1109/ACCESS.2019.2946594.
[15] G. Norris, A. Brookes, and D. Dowell, "The Psychology of Internet Fraud Victimisation: a Systematic Review," *J Police Crim Psychol*, vol. 34, no. 3, pp. 231–245, 2019, doi: 10.1007/s11896-019-09334-5.
[16] H. Xu, L. Shu, P. S. Yu, and B. Liu, "Understanding Pre-trained BERT for Aspect-based Sentiment Analysis," *COLING 2020 - 28th International Conference on Computational Linguistics, Proceedings of the Conference*, pp. 244–250, 2020, doi: 10.18653/v1/2020.coling-main.21.
[17] A. Rogers, O. Kovaleva, and A. Rumshisky, "A primer in bertology: What we know about how bert works," *Trans Assoc Comput Linguist*, vol. 8, pp. 842–866, 2020, doi: 10.1162/tacl_a_00349.
[18] O. Kovaleva, A. Romanov, A. Rogers, and A. Rumshisky, "Revealing the dark secrets of Bert," *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, no. 2018, pp. 4365–4374, 2019, doi: 10.18653/v1/d19-1445.

.