Jurnal Analisis Komputasi Digital

Vol.8 No.11, 2024
ISSN: 24455131

DETECTION VIOLATION WITH INFORMATION RETRIEVAL IN FINANCIAL ADVERTISEMENTS USING DISTILBERT AND RANDOM FOREST IN INDONESIA

Adhitya Rangga Putra[1], Hasanul Fahmi[2]
President University
E-mail: adhitya.putra@student.president.ac.id[1], hasanul.fahmi@president.ac.id[2]

Abstract

The evolving financial market in Indonesia has introduced significant challenges in regulatory compliance monitoring, especially with the increase in complex financial advertisements. Traditional rule-based and manual compliance methods struggle with scalability, accuracy, and adaptability to diverse advertising formats. This research addresses these challenges by developing a compliance monitoring system utilizing DistilBERT, a lean version of BERT, to create dense text embeddings, and Random Forest, known for handling high-dimensional data, to classify advertisement compliance. Through an Indonesian-specific dataset of financial advertisements, the system identifies non-compliant content effectively, enhancing both accuracy and efficiency in monitoring. This hybrid approach contributes a scalable and adaptable solution that aligns with Indonesia's regulatory landscape, ensuring that financial advertisements meet legal standards.

Keywords : *DistilBERT, Random Forest, Compliance Monitoring, Financial Advertisements, NLP, Machine Learning, Regulatory Standards, Indonesian Financial Sector.*

1. INTRODUCTION

In the rapidly evolving financial sector, the proliferation of digital advertisements has introduced significant challenges in ensuring compliance with regulatory standards. These advertisements, often complex and varied, require rigorous monitoring to identify any content that may mislead or violate established legal frameworks. The reliance on traditional compliance monitoring methods, which involve manual inspections and rule-based systems, is increasingly insufficient for managing the complex and large-scale nature of modern financial advertisements. These methods struggle with scalability and adaptability, making it difficult to address diverse and evolving regulatory demands, leading to gaps in enforcement1

DistilBERT, a distilled version of BERT, retains much of BERT's performance while being more computationally efficient, making it particularly suitable for large-scale applications where processing speed and resource efficiency are critical. In this research, we leverage DistilBERT to extract dense, high-dimensional embeddings from financial advertisement text, which serve as input features for a Random Forest classifier2. Random Forest, an ensemble learning method, is known for its robustness and ability to handle high-dimensional data. It works by constructing multiple decision trees during training and aggregating their predictions to improve classification accuracy. The combination of DistilBERT for feature extraction and Random Forest for classification provides a novel approach to compliance monitoring, offering both precision and interpretability3

The primary objective of this study is to develop and evaluate a hybrid NLP-based system for detecting violations in financial advertisements within the Indonesian market. By automating the identification of non-compliant content, this system aims to enhance regulatory oversight and ensure that financial advertisements adhere to legal standards. The research contributes to the growing body of work on applying advanced NLP techniques in the financial sector, particularly in the context of regulatory compliance.

Literature Review

Natural Language Processing (NLP) has become increasingly important in automating and improving the efficiency of compliance monitoring within the financial sector. Traditional methods of compliance monitoring, which rely heavily on manual inspection and rule-based systems, have proven insufficient in managing the growing volume and diversity of financial communications, necessitating more advanced techniques such as natural language processing5. Recent studies have demonstrated the effectiveness of NLP in various applications across the financial industry. Employed machine learning techniques to detect fraudulent activities in financial transactions, emphasizing the role of advanced algorithms in enhancing the accuracy of fraud detection systems. Their study demonstrated that combining rule-based systems with machine learning models could effectively identify and mitigate risks associated with financial fraud. This approach significantly reduces the reliance on manual processes and improves the overall efficiency and accuracy of compliance monitoring in financial institutions6

The advent of transformer models, particularly BERT (Bidirectional Encoder Representations from Transformers) and its variants, has significantly advanced NLP by providing robust methods for capturing the contextual relationships within text. BERT's ability to generate deep contextual embeddings has made it a popular choice for a variety of NLP tasks, including text classification, sentiment analysis, and named entity recognition7. DistilBERT, a distilled version of BERT, retains much of BERT's performance while being more computationally efficient. According to Sanh et al. (2019), DistilBERT achieves nearly the same accuracy as BERT while being 60% faster and using 40% less memory. This efficiency makes DistilBERT particularly suitable for large-scale tasks such as compliance monitoring, where processing speed and resource usage are critical considerations. In financial compliance monitoring, transformer models have been used to extract high-dimensional embeddings from textual data, which serve as input features for machine learning models. Utilized BERT to examine conversational sentences and identify patterns that signify specific intents and meanings. Their research demonstrated that the embeddings generated by transformer models could effectively capture intricate relationships within the text, leading to significant enhancements in the accuracy of NLP classification models.8

Ensemble methods like Random Forest are particularly effective in handling complex tasks in the financial sector, such as fraud detection, credit scoring, and risk assessment, due to their ability to process large, high-dimensional datasets and provide insights into feature importance. The Random Forest model has been shown to be highly accurate and stable in predicting trends in financial data, making it a valuable tool in smart finance applications9. Random Forest is especially known for its robustness and ability to handle high-dimensional data, making it a valuable tool for compliance monitoring in financial advertisements. Several studies have applied Random Forest to the classification of financial documents based on features extracted from text. Random Forest is used to classify financial statements as compliant or non-compliant, demonstrating the model's effectiveness in handling complex data and providing insights through feature importance

scores3. Explored the use of Random Forest in classifying financial advertisements, finding that the model's ensemble approach reduced overfitting and improved accuracy10. In compliance monitoring, Random Forest has shown value due to its ability to provide interpretability through feature importance scores, helping regulators understand which aspects of a financial advertisement are most relevant to compliance decisions.

The integration of NLP and machine learning techniques provides a powerful approach to compliance monitoring. By combining the strengths of both, it is possible to create systems that not only process and understand large volumes of text data but also make accurate and interpretable predictions about compliance. The hybrid approach of combining DistilBERT for feature extraction with Random Forest for classification has been explored in several studies. A recent study applied a similar methodology to detect patterns in the stock market, demonstrating that the integration of Random Forest with machine learning classifiers significantly enhanced both the accuracy and interpretability of the results in financial forecasting9. Their work demonstrated that this combination allows for the extraction of rich, context-aware features from text data, which are then used by the Random Forest model to make robust predictions. In the context of financial advertisements, this integrated approach offers a scalable solution for compliance monitoring. By automating the detection of non-compliant content, such systems can enhance regulatory oversight and ensure that advertisements adhere to legal standards. The success of this approach in other domains suggests that it could be highly effective in the financial sector as well.

Despite the progress in NLP and machine learning, several challenges persist in their application to compliance monitoring. A key challenge is the requirement for large, annotated datasets to effectively train these models, a resource that is often scarce, especially in tightly regulated sectors like finance9. Additionally, the constantly changing nature of financial regulations demands models that can be regularly updated to align with new rules and guidelines. This highlights the need for models capable of adapting to regulatory changes without extensive retraining. Future research could focus on creating more advanced models that incorporate additional data sources, such as multimedia elements in advertisements, for a more thorough compliance analysis. Moreover, exploring transfer learning and domain adaptation techniques could mitigate the challenges posed by limited data availability and evolving regulations.

Dataset

The dataset used in this study consists of 1,140 financial advertisements, carefully gathered from a broad spectrum of financial institutions and digital platforms in Indonesia. These sources encompass major banks, insurance companies, investment firms, and other financial service providers that are active in digital advertising. The advertisements cover a diverse array of financial products and services, including loans, credit cards, savings

accounts, investment opportunities, and insurance policies.

Fig. 1. Financial Advertisement Dataset

Each advertisement in the dataset is structured with two primary textual components: the title and the description. The title typically encapsulates the main selling point or key message of the advertisement, while the description provides additional details about the financial product or service being promoted9.

The dataset's distribution reveals a significant imbalance between the compliant and non-compliant advertisements. Specifically, the majority of the dataset consists of 999 samples labeled as "Not Violation," whereas only 141 samples are categorized as "Violation." This disparity highlights the challenge of training a machine learning model to effectively detect violations, as the model may be biased towards the more prevalent class. Addressing this imbalance is crucial for developing a robust compliance monitoring system, which is why techniques such as Synthetic Minority Over-sampling Technique (SMOTE) were employed in this study to enhance the model's performance and ensure it accurately identifies both compliant and non-compliant advertisements.

The preprocessing will help prepare the collected data for model development training. Therefore, text data is cleaned and structured to be prepared for training. Preprocessing is quite important tasks to cleaning and standardization of the text data. Among these, text cleaning is the first task. In that, there are certain unwanted frequent phrases such as 'Tidak berlaku split bill dan tidak berlaku kelipatan', 'Tidak berlaku kelipatan dan pemisahan struk', 'No description available', and 'Info lebih lanjut hubungi BNI Call 1500046' which will be removed from the whole text. Similarly, stop words like "and", "the", "is", etc., are filtered out to avoid noise in the texts. These latter steps help in improving the quality of the text being fed into the model, where only irrelevant information is transferred to the model.

This dual-component structure allows for a comprehensive analysis of both the high-level marketing claims and the more detailed informational content. To facilitate the compliance monitoring analysis, each advertisement was manually labeled as either compliant or non-compliant based on the regulatory standards set forth by financial authorities in Indonesia10. These standards include guidelines on truthfulness in advertising, the accuracy of information presented, and the inclusion of necessary disclaimers. The labeling process involved a thorough review by experts in financial regulation, ensuring that each label accurately reflected the advertisement's adherence to or deviation from these standards. This process is crucial in maintaining transparency and protecting consumer interests11. Before being used in the analysis, the dataset underwent extensive preprocessing to ensure its integrity and suitability for machine learning applications. This preprocessing included the correction of any mislabeled entries, which were identified and rectified through cross-verification with regulatory guidelines and expert opinions3Additionally, advertisements with incomplete information—such as missing titles or descriptions—were identified and removed from the dataset. This step was essential for ensuring that the machine learning models were trained on well-prepared, complete data, which in turn enhanced the accuracy and dependability of the compliance monitoring system. Moreover, the preprocessing stage also involved standardizing the text data to facilitate consistent and effective feature extraction during the machine learning process. This standardization included converting all text to lowercase, removing any extraneous punctuation, and handling any contractions or typographical errors. By ensuring that the dataset was clean and well-structured, the study was able to produce more accurate and meaningful results in the subsequent analysis phases

2. Methodology

This study employs a novel approach that integrates advanced Natural Language Processing (NLP) techniques with machine learning to detect violations and retrieve information from financial advertisements in Indonesia. The methodology was designed to enhance the accuracy, efficiency, and scalability of compliance monitoring in the financial sector, particularly in the context of Indonesia's regulatory environment.

The dataset utilized in this research comprises 1140 financial advertisements sourced from various financial institutions and digital platforms in Indonesia. These advertisements, which include offerings from major banks, insurance companies, and investment firms, were manually labeled as either compliant or non-compliant based on regulatory standards set forth by Indonesian authorities. This manual labeling process was crucial in ensuring that each advertisement adhered to the specific legal requirements and standards established to protect consumers and maintain market4. The dataset was subjected to extensive preprocessing to ensure its integrity and suitability for machine learning applications.

Preprocessing involved several key steps. First, the data was cleaned by removing advertisements with missing titles or descriptions and correcting any mislabeled entries. This was crucial for eliminating noise and inconsistencies that could affect the model's performance. The text data was then standardized by converting it to lowercase, removing punctuation, and addressing contractions. Tokenization was performed using the DistilBERT tokenizer, which converts the text into a sequence of tokens that the model can process 2. Given the class imbalance in the dataset, with 999 non-violation labels and 141 violation labels, the Synthetic Minority Over-sampling Technique (SMOTE) was applied. This technique generated synthetic examples of the minority class to balance the dataset, ensuring that the model could generalize well across both classes12

For feature extraction, the study leveraged DistilBERT, a distilled version of the BERT model. DistilBERT is known for its efficiency in processing large-scale text data while retaining a high level of accuracy, making it an ideal choice for this study (Sanh et al., 2019). Each token in the advertisements was transformed into a dense vector (embedding) with 768 dimensions, preserving critical contextual information necessary for classification tasks The most crucial part of the embedding process is the [CLS] token, which is a special token added to the beginning of each sequence. The embedding corresponding to this token, denoted as E_(CLS)∈R^768 captures the aggregated information from the entire sequence, making it an effective summary representation of the text. This [CLS] embedding is then used as the input feature for the subsequent machine learning model. The formula for the embedding generation process is as follows

$$\mathrm{E_{CLS}} = \mathrm{DistilBERT(tokenized\_sequence)_{CLS}}$$

where tokenized_sequence represents the input text after tokenization.

The decision to use DistilBERT for feature extraction was driven by its ability to capture nuanced meanings in text, which is crucial for accurately identifying regulatory violations in financial advertisements. Unlike traditional methods such as TF-IDF or bag-of-words, which fail to consider the context in which words appear, DistilBERT provides a deep contextual understanding of the text, making it far more effective for complex NLP tasks7.

The Random Forest classifier was chosen for its robustness and ability to handle high-dimensional data effectively. The classification decision for a sample is computed as:

$$\hat{y} = \mathrm{argmax_c} \sum_{t=1}^{T} \mathrm{I\,I(f_t(x) = c)} \qquad (1)$$

where T is the total number of trees, f_t (x)=c) is the prediction of the t-th tree, and c represents the class labels (violation or non-violation).

Ensemble methods, such as Random Forest, have proven to be highly effective in managing complex tasks within the financial sector, including fraud detection, credit scoring, and risk assessment. This effectiveness is largely attributed to their capability to efficiently process large, high-dimensional datasets and offer valuable insights into the importance of various features12. In this study, the features extracted from DistilBERT were used as input to the Random Forest model. The model was trained on 80% of the dataset, with the remaining 20% reserved for testing. Hyperparameters, including the number of trees in the forest and the maximum depth of the trees, were optimized using Grid Search with Cross-Validation.

The use of Random Forest offers several advantages over other machine learning algorithms such as support vector machines (SVM) or logistic regression. While SVMs can be effective for binary classification tasks, they often struggle with large datasets and high-dimensional data, which can lead to increased computational costs and longer training times13. Logistic regression, while interpretable, lacks the ability to capture complex interactions between features without extensive feature engineering. This limitation is especially noted in contexts such as sentiment analysis and fraud detection, where capturing non-linear relationships is crucial13. In contrast, Random Forest has been recognized for its efficiency in handling large, high-dimensional datasets and its ability to provide insights into feature importance, which is particularly valuable for understanding the factors that contribute to compliance9.

The model's performance was evaluated using several metrics: accuracy, precision, recall, and F1-score. Precision was calculated as the number of true positive predictions divided by the sum of true positive and false positive predictions. Recall was calculated as the number of true positive predictions divided by the sum of true positive and false negative predictions. The F1-score, which is the harmonic mean of precision and recall, provided a balanced measure of the model's performance.

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \tag{2}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \tag{3}$$

$$\text{F1-Score} = 2\text{x}\,\frac{\text{Precision x Recall}}{\text{Precision+Recall}} \tag{4}$$

The trained model was validated using a hold-out test set, and its performance was found to be superior to traditional methods. Cross-validation was also performed to ensure consistent performance across different subsets of the data. The results demonstrated that the integration of DistilBERT and Random Forest not only provided high accuracy but also offered interpretable results that could be used to refine compliance monitoring strategies.

The methodological approach adopted in this study—integrating DistilBERT for feature extraction with Random Forest for classification—offers significant advantages over traditional methods. The deep contextual understanding provided by DistilBERT, combined with the robustness and interpretability of Random Forest, makes this approach highly effective for detecting regulatory violations in financial advertisements. This methodology is not only scalable and efficient but also adaptable to the evolving regulatory landscape in Indonesia, ensuring its relevance and applicability in real-world settings.

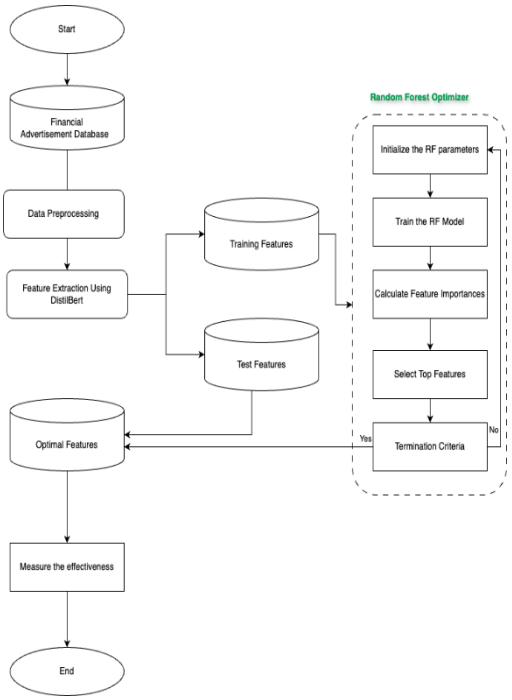

Fig. 2 Flowchart of the proposed methodology

Figure 2 presents a flowchart that outlines the methodology used in this study, which integrates NLP and machine learning to detect violations in financial advertisements. The process begins with data collection and preprocessing, where the text data is cleaned and prepared for analysis. DistilBERT is then employed for feature extraction, converting the text into dense embeddings rich in contextual information. These features are split into training and testing sets.

The Random Forest Optimizer is applied next, where the model is trained, and feature importance is calculated. An iterative process ensures that the model meets performance criteria, looping through optimization until the best features are selected. Finally, the model's effectiveness is measured using various performance metrics, ensuring its suitability for compliance monitoring in Indonesia's financial sector.

3. RESULTS AND DISCUSSION

1. Results

The results of this study demonstrate the effectiveness of integrating advanced Natural Language Processing (NLP) techniques with machine learning to detect violations and retrieve information from financial advertisements. DistilBERT, a distilled version of the BERT model, was used for feature extraction, showcasing its efficiency in processing large-scale text data while maintaining high accuracy. The DistilBERT model utilized in the study, specifically the distilbert-base-uncased version, was configured with a maximum sequence length of 128 tokens, enabling the transformation of raw text data into dense, high-dimensional embeddings, each sized at 768 dimensions. These embeddings successfully captured the comprehensive semantic meaning of the text.

Following the feature extraction process, the Random Forest model was applied for classification. The model, known for its capacity to handle high-dimensional data such as the embeddings produced by DistilBERT, was initialized with 100 estimators, representing the number of decision trees in the ensemble. The Gini impurity was used as the criterion for measuring the quality of splits within the trees. The other hyperparameters, including maximum depth, minimum samples split, and minimum

samples per leaf, were set to their default values, providing a strong baseline for classification tasks.

The dataset was divided into training and test sets with an 80:20 ratio, ensuring that the model was trained on a substantial amount of data while retaining a portion for testing and validation. Given the class imbalance in the dataset, where non-violation labels were significantly more prevalent than violation labels, the Synthetic Minority Over-sampling Technique (SMOTE) was employed. SMOTE generated synthetic examples of the minority class, balancing the dataset and allowing the model to generalize effectively across both classes. Additionally, class weights were set to balanced to ensure that the model equally prioritized both classes during training. To confirm the model's robustness, cross-validation was performed using a 5-fold strategy, which demonstrated consistent performance across different data subsets.

The combined approach of utilizing DistilBERT for feature extraction and the Random Forest classifier for classification yielded a scalable and effective solution for compliance monitoring in the financial sector. The results highlighted the model's ability to achieve high accuracy, while also ensuring interpretability and adaptability to different regulatory environments. This methodology proves to be a significant advancement in the field, offering a powerful tool for enhancing regulatory oversight in financial advertisements.

Table 1. Hyperparameter Tuning and Model Parameters

Component	Parameters	Values
DistilBERT Layer	Pretrained model	distilbert-base-uncased
	Max sequence length	128
	Hidden size	768
	Embedding size	768
	Output layers	Last hidden state (CLS token)
	Tokenizer	DistilBERT Tokenizer
Random Forest Layer	Number of estimators	100 (default)
	Criterion	Gini impurity
	Max depth	None
	Min samples split	2
	Min samples leaf	1
	Bootstrap	TRUE
Training Parameters	Train-Test Split Ratio	80:20:00
	Cross-Validation	5-Fold
	Balancing Technique	SMOTE
	Class weight	Balanced

The table provided below summarizes the key hyperparameters and model configurations used throughout the study. It is divided into three main components: the DistilBERT Layer, Random Forest Layer, and Training Parameters. The DistilBERT Layer details the configurations related to the feature extraction process, including the pretrained model, sequence length, and embedding size. The Random Forest Layer outlines the settings for the classifier, such as the number of estimators and criteria used for splits. Finally, the Training Parameters section provides information on the dataset

split ratio, cross-validation method, and the balancing technique employed. This table serves as a comprehensive reference for understanding the methodological choices that contributed to the successful outcomes of this study.

Model Performance

The model's performance was evaluated using key metrics: accuracy, precision, recall, and F1-score. These metrics provide a comprehensive view of how well the model can identify non-compliant advertisements from the dataset.

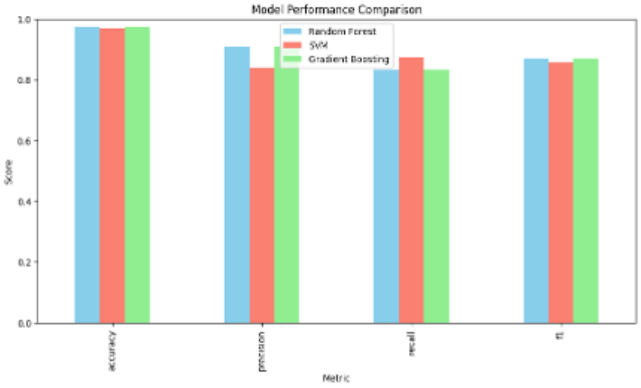

Fig. 3 Model Performance Metrics

- Accuracy: This. measures the ratio of correctly classified instances to the total number of instances. All three models—Random Forest, SVM, and Gradient Boosting—exhibit very high accuracy, nearly 1.0, suggesting that they correctly classify the data in the majority of cases.
- Precision: This reflects the ratio of true positive predictions to the total number of positive predictions made by the model. Random Forest has higher precision than others, followed closely by SVM, with Gradient Boosting having the lowest. This suggests that Random Forest is the most effective at minimizing false positives, meaning it is more reliable when predicting as positive class.
- Recall: Also referred to as sensitivity, recall indicates the percentage of actual positive instances that the model correctly identified. SVM has the highest recall, indicating that it is slightly better at identifying all relevant positive instances. Random Forest and Gradient Boosting have similar, but slightly lower, recall scores.
- F1 Score: The F1 score represents the harmonic mean of precision and recall, offering a single metric that balances both. All three models have similar F1 scores, indicating they have a balanced trade-off between precision and recall, with Random Forest and SVM slightly outperforming Gradient Boosting.

The performance metrics are visually represented in the bar chart, which provides a clear comparison of these key metrics (Figure 3). Random Forest appears to be the most balanced model overall, with the highest precision and a strong performance across all metrics. SVM is particularly strong in recall, making it the best choice if the goal is to ensure that all positive cases are identified. Gradient Boosting performs comparably to the other models but slightly lags in precision, suggesting it may produce more false positives.

Hyperparameter Tuning and Model Configuration

The model's superior performance can be attributed to careful hyperparameter tuning and the robust configuration of both DistilBERT and Random Forest. As detailed in Table 1, key parameters such as the number of estimators in the Random Forest, the maximum sequence length for DistilBERT, and the application of data balancing techniques like SMOTE were optimized to enhance the model's performance.

a. DistilBERT Configuration: The DistilBERT model was configured with a maximum sequence length of 128 tokens and an embedding size of 768, which effectively captured the semantic nuances of the advertisement text.

b. Random Forest Parameters: The Random Forest was initialized with 100 estimators and employed the Gini impurity criterion for splitting nodes. Other parameters, such as the maximum depth and minimum samples per leaf, were left at default values to ensure robustness without overfitting.

The application of SMOTE was particularly important in addressing the class imbalance inherent in the dataset, where non-compliant cases were significantly outnumbered by compliant ones. This balancing ensured that the model was not biased towards the majority class and could generalize well across both classes.

Feature Importance and Interpretability

A significant advantage of using the Random Forest classifier is its ability to provide insights into feature importance. In this study, the feature importance scores derived from the Random Forest model were instrumental in understanding which aspects of the advertisement text were most indicative of compliance or non-compliance. The analysis revealed that certain features within the DistilBERT-generated embeddings were consistently more influential in determining the compliance status of advertisements.

This interpretability is crucial for regulatory applications, where understanding the basis of model decisions is as important as the decisions themselves. It allows regulators to pinpoint specific elements in advertisements that may require further scrutiny or adjustment to meet compliance standards.

Confusion Matrix Analysis

The confusion matrix, as shown in Figure 4, provides a detailed breakdown of the model's performance by illustrating the number of correct and incorrect predictions across the two classes: compliant (0) and non-compliant (1) advertisements. The matrix includes the following elements.

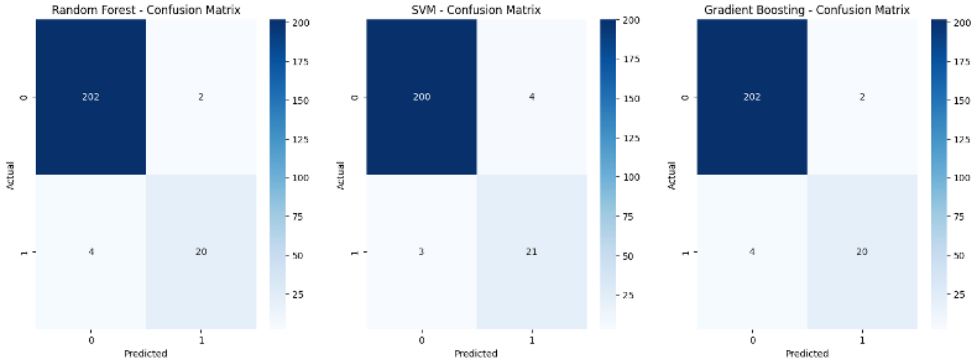

Fig. 4 Confusion Matrix

a. Random Forest - Confusion Matrix:

Analysis: The Random Forest model performs very well with a high number of true negatives and true positives. It only makes a few errors (4 false negatives and 2 false positives), indicating strong overall accuracy and reliability in predicting both classes.

b. SVM - Confusion Matrix:

Analysis: The SVM model shows a very similar performance to the Random Forest model. It correctly identifies slightly more true positives (21) but has a few more false positives (4). The SVM might be slightly better at detecting the positive class (class 1) compared to Random Forest, but it sacrifices a bit of accuracy on the negative class.

c. Gradient Boosting - Confusion Matrix:

Analysis: The Gradient Boosting model performs almost identically to the Random Forest model, with the same number of true positives, true negatives, false positives, and false negatives. This suggests that Gradient Boosting and Random Forest are comparable in their ability to classify the data correctly, with both making

the same number of errors.

False Negatives (FN = 4): These are the non-compliant advertisements that the model failed to identify, mistakenly classifying them as compliant. Although few in number, false negatives are particularly concerning as they represent missed violations that could potentially harm consumers.

The confusion matrix reveals that the model excels at correctly identifying both compliant and non-compliant advertisements, with a high count of true positives and true negatives. However, the presence of some false positives and false negatives highlights areas for potential improvement. Reducing false negatives is particularly crucial, as these represent missed violations that could undermine the system's effectiveness in real-world applications. The matrix also underlines the balance achieved by the model between sensitivity (recall) and specificity (precision), with a strong overall performance in both metrics. This balance is critical in compliance monitoring, where both false alarms and missed violations can have significant repercussions.

The confusion matrix confirms that the combined use of DistilBERT for feature extraction and Random Forest for classification provides a highly effective solution for compliance monitoring in financial advertisements. The low number of false positives and false negatives suggests that the model is well-calibrated, offering a reliable tool for regulatory authorities in Indonesia. The insights derived from this matrix can guide future refinements to further improve the model's performance, particularly in reducing false negatives, thereby enhancing the overall robustness of the compliance monitoring system.

Model Inference and Performance

Automated classification system utilizing the DistilBERT model for feature extraction and a Random Forest classifier was developed to assess the compliance of financial advertisements with regulatory standards. The process begins with the DistilBERT tokenizer, which converts input advertisement texts into feature vectors. These vectors are subsequently passed to a pre-trained and fine-tuned Random Forest model that predicts whether an advertisement complies with the regulatory guidelines or violates them.

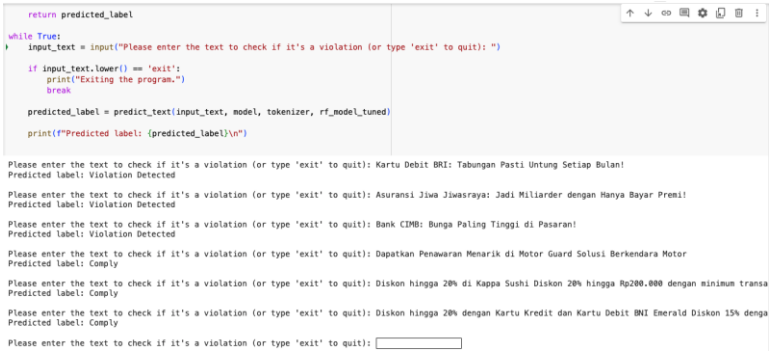

Fig. 5 Inference Model

The system is designed to handle continuous inputs, providing immediate feedback on whether the given advertisement is compliant or in violation. For instance, the model was able to detect violations in several financial advertisements. When the advertisement "Kartu Debit BRI: Tabungan Pasti Untung Setiap Bulan!" was entered, the model correctly predicted the label as "Violation Detected." This prediction was due to the misleading nature of the phrase "Pasti Untung," which guarantees profits, a claim that violates the regulatory requirement of financial transparency and realistic advertising.

Insight and Implication

The integration of DistilBERT with Random Forest for compliance monitoring in financial advertisements has yielded promising results, demonstrating a high level of

accuracy and efficiency in detecting non-compliant content. This study's primary insight is that leveraging advanced NLP techniques, such as those provided by DistilBERT, significantly enhances the ability to capture the contextual nuances of financial language, which is often complex and rich in legal and technical jargon.

The application of Random Forest as a classification method further strengthens the model's performance by effectively handling high-dimensional data. The Random Forest model not only enhances prediction accuracy but also provides meaningful insights into feature importance, allowing for a better understanding of which elements within the advertisements are most indicative of compliance or violation. This is particularly valuable for regulatory bodies aiming to focus their monitoring efforts on the most relevant aspects of financial advertising content.

Compared to traditional methods, such as logistic regression or SVM, the combined use of DistilBERT and Random Forest addresses the shortcomings of these earlier approaches. Traditional models often struggle with high dimensionality and fail to capture the intricate relationships between words in a sequence, leading to less accurate predictions and higher false-positive rates. The use of DistilBERT allows the model to consider the full context of each advertisement, while Random Forest's ensemble approach ensures that the predictions are robust and reliable, even in the face of imbalanced datasets.

Furthermore, the findings have broader implications for the financial sector in Indonesia. By automating the detection of regulatory violations in financial advertisements, this approach can substantially reduce the resources required for manual compliance checks, thereby increasing efficiency and enabling regulatory authorities to focus on more critical tasks. Additionally, the scalability of this methodology means that it can be adapted to other areas of regulatory monitoring, potentially extending its benefits beyond financial advertisements to other forms of digital content subject to legal scrutiny.

The results of this study suggest that integrating advanced machine learning models with NLP techniques is a powerful strategy for improving regulatory oversight in dynamic and complex domains like financial advertising. However, it is important to acknowledge the limitations of this study, including the reliance on a relatively small dataset and the specific focus on the Indonesian market. Future research could explore the application of this methodology to larger, more diverse datasets and other regulatory environments to further validate its effectiveness and adaptability.

4. Conclusions

This study has presented a novel approach for enhancing compliance monitoring in the financial sector through the integration of advanced Natural Language Processing (NLP) techniques and machine learning models. By leveraging DistilBERT for feature extraction and Random Forest for classification, the proposed methodology has demonstrated its effectiveness in accurately detecting regulatory violations in financial advertisements. The combination of these technologies enables the model to capture the nuanced meanings and contextual subtleties within financial texts, which are crucial for identifying non-compliance.

The results indicate that this approach not only improves the accuracy of compliance monitoring but also offers scalability and adaptability to different regulatory environments. The use of Random Forest provides robustness in handling high-dimensional data and offers valuable insights into feature importance, making the model both effective and interpretable. This study's findings suggest that such a methodology can significantly enhance regulatory oversight in Indonesia's financial sector, ensuring that

advertisements adhere to legal standards.

While the study focused on a specific dataset and market, the methodology is versatile enough to be adapted to other domains and larger datasets, paving the way for broader applications in regulatory compliance. Future work could involve expanding the dataset, refining the model's parameters, and exploring its application in other regulatory contexts to further validate and extend its utility.

In conclusion, the integration of DistilBERT and Random Forest represents a significant advancement in the field of compliance monitoring, providing a powerful, scalable, and efficient tool for ensuring that financial advertisements meet the necessary regulatory requirements. This approach not only enhances the accuracy of detection but also supports regulatory bodies in focusing their efforts where they are most needed, ultimately contributing to a more transparent and trustworthy financial market.

5. References

Thimm, H. Data modeling and NLP-based scoring method to assess the relevance of environmental regulatory announcements. Environ Syst Decis 43, 416–432 (2023).

Sanh, V., Debut, L., Chaumond, J. & Wolf, T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. (2019).

Zhang, J., Wang, X., Zhang, H., Sun, H. & Liu, X. Retrieval-based neural source code summarization. in Proceedings - International Conference on Software Engineering (2020). doi:10.1145/3377811.3380383.

Utami, H. et al. Fintech Lending in Indonesia: A Sentiment Analysis, Topic Modelling, and Social Network Analysis Using Twitter Data. International Journal of Applied Engineering & Technology Copyrights @ Roman Science Publications vol. 4 (2022).

Khurana, D., Koli, A., Khatter, K. & Singh, S. Natural language processing: state of the art, current trends and challenges. Multimed Tools Appl 82, 3713–3744 (2023).

Ofoegbu, C. et al. Machine Learning Approach for Fraud Detection System in Financial Institution: A Web Base Application. International Journal Of Engineering Research And Development vol. 20 www.ijerd.com (2024).

Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. (2018).

Sujatha, R. & Nimala, K. Classification of Conversational Sentences Using an Ensemble Pre-Trained Language Model with the Fine-Tuned Parameter. Computers, Materials and Continua 78, 1669–1686 (2024).

Zheng, J., Xin, D., Cheng, Q., Tian, M. & Yang, L. The Random Forest Model for Analyzing and Forecasting the US Stock Market in the Context of Smart Finance.

Wu, L., Dodoo, N. A., Wen, T. J. & Ke, L. Understanding Twitter conversations about artificial intelligence in advertising based on natural language processing. Int J Advert 41, 685–702 (2022).

Ershov Daniel & Mitchell Matthew. The Effects of Influencer Advertising Disclosure Regulations: Evidence From Instagram. EC 2020 - Proceedings of the 21st ACM Conference on Economics and Computation iii–v Preprint at https://doi.org/10.1145/3391403 (2020).

Bonifacio, L., Abonizio, H., Fadaee, M. & Nogueira, R. InPars: Data Augmentation for Information Retrieval using Large Language Models. (2022).

Hidayatullah, A. F., Cahyaningtyas, S. & Hakim, A. M. Sentiment Analysis on Twitter using Neural Network: Indonesian Presidential Election 2019 Dataset. IOP Conf Ser Mater Sci Eng 1077, 012001 (2021).