USING ARTIFICIAL INTELLIGENCE TO COMBAT SOCIAL MEDIA SCAMS IN THAILAND

by

Voravit Cheepphitaksakul, M.Sc.(IT)

DISSERTATION

Presented to the Swiss School of Business and Management Geneva

In Partial Fulfillment

Of the Requirements

For the Degree

DOCTOR OF BUSINESS ADMINISTRATION

SWISS SCHOOL OF BUSINESS AND MANAGEMENT GENEVA

<MONTH OF GRADUATION, YEAR>

USING ARTIFICIAL INTELLIGENCE TO COMBAT SOCIAL MEDIA SCAMS IN THAILAND

by

Voravit Cheepphitaksakul

APPROVED BY

Dissertation chair - Dr. Gualdino Cardoso

RECEIVED/APPROVED BY:

Renee Goldstein Osmic

Admissions Director

Dedication

Dedicated to my parents and my friends for all the help and support they have provided me during this journey.

Acknowledgements

First and foremost I would like to thank my mentor for his support through not only my dissertation but my time at SSBM. I have been fortunate enough to work with a few of you extensively and feel grateful for all that I have learned from you. While I am grateful to my entire committee, I must pay particular thanks to my mentor Dr. Mario Silic for you have spent many for the advice that you have provided to shape my thesis paper. I appreciate all the time you have spent working through ideas and reading endless drafts of proposals, papers, and this thesis.

Thanks to my friends and family members. Personal and professional support has proven to be my strength through this journey. You each have offered me support and friendship at some point during my journey which I found invaluable.

ABSTRACT

USING ARTIFICIAL INTELLIGENCE TO COMBAT SOCIAL MEDIA SCAMS IN THAILAND

Voravit Cheepphitaksakul
2026

Dissertation Chair: Dr.Mario Silic

Social media scams have become a major threat in Thailand, targeting individuals and businesses through fraudulent messages, fake advertisements, and impersonation schemes. Traditional scam detection methods struggle to keep up with the rapid evolution of these deceptive tactics. This research proposes an AI-driven approach to detect and prevent social media scams in real-time. We develop a Thai-language scam detection model using WangchanBERTa, a BERT-based NLP model trained on a labeled dataset of scam-related content. The model is deployed as an API using FastAPI via A web dashboard. This AI-powered solution enhances cybersecurity efforts, providing an efficient and scalable defense against social media scams in Thailand

TABLE OF CONTENTS

CHAPTER I:
INTRODUCTION

1.1 Introduction

The rapid expansion of digital communication has reshaped the social and economic landscape of Thailand, creating unprecedented opportunities for interaction while simultaneously exposing individuals and businesses to new forms of cyber risk. Among these emerging threats, social media scams have become one of the most pervasive and damaging. These scams exploit the openness, speed, and anonymity of online platforms, enabling fraudsters to target large populations with minimal effort.

In recent years, Thailand has witnessed a dramatic escalation in online fraud cases. Reports from the Electronic Transactions Development Agency (ETDA) indicate that more than **276,000 online scam complaints were filed in 2023**, with total estimated financial losses exceeding **40 billion THB**, marking one of the highest annual increases in the nation's digital history (ETDA, 2024). Similarly, data from the Cyber Crime Investigation Bureau (CCIB) highlight that impersonation scams, investment frauds, and fake online marketplaces remain among the most common patterns observed, often involving sophisticated linguistic manipulation rather than purely technical intrusion (CCIB, 2024).

The linguistic nature of these scams adds a layer of complexity to fraud detection efforts. Thai language structure—characterized by the absence of spaces between words, diverse grammatical patterns, and frequent use of informal particles—poses difficulties

for rule-based or keyword-based detection methods. Furthermore, the prevalence of Thai-English code-switching, phonetic spellings, and emojis on social media platforms complicates the detection process. As scammers continually refine their communication strategies to avoid detection, traditional approaches struggle to keep pace.

These challenges indicate a critical need for more adaptive, intelligent, and context-aware fraud detection mechanisms. With recent developments in Natural Language Processing (NLP) and transformer-based models, new opportunities have emerged to analyze text patterns more deeply and accurately. Models such as BERT (Devlin et al., 2019) and its Thai-language adaptations like WangchanBERTa (Lowphansirikul et al., 2021) demonstrate the potential for understanding semantic nuances in ways traditional algorithms cannot.

This chapter sets the foundation for a dissertation that explores how NLP techniques can be leveraged to combat social media scams in Thailand. The study aims to develop, train, and evaluate a Thai-language scam detection model using transformer-based architecture and assess its potential for deployment in real-world environments.

1.2 Background of Social Media Scams in Thailand

Social media scams are not simply technical problems; they are communication-based crimes that exploit trust, urgency, and psychological pressure. Fraudsters rely on scripted narratives, fabricated identities, and persuasive language tailored to the online behaviors of Thai users. According to the Bank of Thailand (2024), the most common scam types include:

Impersonation scams (posing as bank officers or government agencies)

Loan and financial assistance scams

Investment scams, especially high-yield crypto schemes

Fake parcel delivery and customs fee scams

Job and part-time income scams

These scams often follow recognizable linguistic and emotional patterns—such as politeness markers, expressions of urgency, appeals to authority, or emotionally charged narratives—making them suitable candidates for NLP-based analysis (Rehman et al., 2022).

The rise of social engineering as the dominant vector for cybercrime aligns with global trends. Ferrara (2020) notes that online deception increasingly relies on conversational persuasion and mimicry rather than malware. This shift underscores the need for computational methods capable of detecting subtle linguistic cues embedded within ordinary conversations.

In Thailand, the proliferation of LINE groups, Facebook pages, and e-commerce communities has further accelerated the speed at which scams spread. Closed-group communication and forwarding features allow fraudulent content to circulate widely before detection, often only after significant harm has occurred. This environment makes reactive, manual reporting systems insufficient.

The combination of linguistic complexity, cultural communication norms, and platform dynamics creates a unique environment where advanced NLP models are not only beneficial but necessary.

1.3 Problem Statement

Despite significant advancements in Thailand's digital economy, the country continues to face an escalating wave of social media scams that traditional detection approaches cannot adequately address. Several institutions—including ETDA, CCIB, and the Bank of Thailand—have highlighted that existing fraud detection mechanisms rely heavily on manual review, fixed keyword rules, or user reports submitted after financial losses have already occurred (ETDA, 2024; CCIB, 2024). Such reactive systems lack the ability to interpret evolving scam scripts, subtle linguistic cues, or the informal communication patterns characteristic of Thai social media platforms.

Furthermore, many scam messages operate through dynamic psychological manipulation rather than technical exploitation. They contain culturally embedded politeness particles, blended English–Thai expressions, or persuasive narratives framed to mimic legitimate communication. These linguistic complexities render many traditional filtering systems ineffective, as they are unable to process semantic nuances or contextual variations.

In addition, there is a clear research gap in applying modern NLP methods—specifically transformer-based models—to Thai-language scam detection. While global studies have explored NLP applications for fraud detection in English and other widely resourced languages (Adewumi & Akinbohun, 2022; Ahmad et al., 2021), there remains limited scholarly work tailored to Thai linguistic structures and online communication behaviors.

This gap limits the development of automated tools capable of supporting Thai users and institutions against increasingly sophisticated scams.

Thus, the core problem addressed in this dissertation is the absence of a **robust, Thai-language NLP framework** capable of detecting social media scam messages accurately and efficiently, while accounting for the unique linguistic and cultural attributes of Thai digital communication

1.4 Purpose of Research

The purpose of this research is to design, develop, and evaluate a Thai-language scam detection model using Natural Language Processing (NLP) techniques, specifically leveraging transformer-based architectures such as WangchanBERTa. This study aims to:

Analyze the linguistic characteristics of Thai scam messages, including patterns of persuasion, impersonation, urgency, and emotional manipulation commonly found on social media platforms in Thailand.

Develop and fine-tune a transformer-based NLP model capable of distinguishing between scam and non-scam messages with high accuracy.

Evaluate the model's performance using established machine learning metrics, enabling comparison with traditional detection approaches.

Deploy the model as a functional API in a real-time environment, demonstrating its practical applicability for implementation by digital service providers, financial institutions, or public safety agencies.

By achieving these objectives, the research aims to produce an evidence-based, operational NLP tool that can contribute to Thailand's broader efforts in cybercrime prevention and digital safety.

1.5 Significance of the Study

Academic Significance

This study contributes to the growing body of knowledge on computational linguistics, cyber fraud, and NLP for low-resource languages. While the global community has made substantial progress in transformer-based NLP research, Thai-language applications—particularly in cybersecurity—remain underexplored. This dissertation fills that gap by providing:

- A comprehensive linguistic analysis of Thai scam messages
- A reproducible NLP pipeline tailored to Thai text
- An empirical evaluation of transformer-based performance in fraud detection
- Insights for future cross-lingual or multilingual scam detection research

Practical Significance

The practical implications of this study are substantial:

- **Early Scam Detection:**

  An NLP-based system can help organizations flag deceptive messages before victims engage with fraudsters.

- **Financial and Social Protection:**

  With online scams costing Thailand billions of baht annually, an automated detection tool can reduce losses and protect vulnerable groups.

- **Support for Thai Law Enforcement:**

  Agencies such as CCIB may use NLP tools to analyze large volumes of suspicious messages or identify scam networks more efficiently.

- **Improved Platform Safety:**

  Social media companies, banks, and e-commerce providers can integrate the model to enhance user trust and reduce harmful interactions.

- **Public Awareness:**

  Insights from the linguistic patterns identified may support awareness campaigns, educational programs, and digital literacy efforts.

Overall, this research strengthens Thailand's digital resilience by demonstrating how AI can support cybersecurity in a culturally and linguistically informed manner.

1.6 Research Questions

1. What types of social media scams are prevalent in Thailand?
2. Can AI detect scam messages in Thai with high accuracy?
3. How effective is an AI model when deployed in real-time environments?

1.7 Chapter Summary

This chapter introduced the research topic, highlighting the scale and urgency of social media scams in Thailand, the linguistic complexities involved, and the limitations of existing detection mechanisms. It outlined the core problem driving the study, articulated its purpose, and emphasized the significance of developing a Thai-language NLP-based scam detection model. The chapter also presented the research questions guiding the dissertation.

The next chapter provides a detailed review of existing literature on NLP technologies, transformer models, cyber fraud detection, and sociolinguistic perspectives relevant to online deception.

CHAPTER II:
REVIEW OF LITERATURE

2.1 Theoretical Framework

The theoretical foundation for this study is built upon interdisciplinary concepts from natural language processing, artificial intelligence, and cybercrime communication. The framework positions social media scams as *linguistic and behavioral phenomena* that can be analyzed through computational methods. Unlike technical cyberattacks, which rely on system vulnerabilities, social media scams primarily exploit linguistic cues, emotional triggers, and persuasive communication strategies (Rehman et al., 2022).

This study adopts a theoretical view that scam messages are shaped by identifiable patterns—such as impersonation language, urgency framing, and psychological appeal—that can be systematically captured using NLP models. Ferrara (2020) emphasizes that modern online deception increasingly relies on subtle manipulation rather than overt malicious code, suggesting that linguistic patterns offer a reliable pathway for detection. At the same time, the transformer architecture (Vaswani et al., 2017), which underpins modern NLP, provides a theoretical scaffold for understanding how models can capture contextual meaning in text. Unlike earlier methods that relied on bag-of-words or static embeddings, transformers interpret meaning bidirectionally, making them suitable for nuanced languages like Thai.

Thus, the theoretical framework in this chapter integrates:

1. **Linguistic deception theories** – identifying how scammers craft persuasive messages.
2. **NLP semantics and contextual modeling** – enabling the detection of subtle textual cues.
3. **Cyber fraud behavioural models** – explaining how victims respond to scam communication.

Together, these theories create a foundation for investigating how NLP can be applied to enhance scam detection in Thailand's social media environment.

2.2 Understanding Natural Language Processing (NLP)

Natural Language Processing (NLP) is a field of artificial intelligence that enables computers to understand, interpret, and generate human language. Its development has significantly expanded the capabilities of digital systems to perform tasks such as classification, translation, question-answering, and content moderation. The relevance of NLP to scam detection lies in its ability to analyze text-based communication at scale, identifying patterns that humans may not readily notice.

As digital interactions become increasingly text-driven, NLP has emerged as a critical technology in cybersecurity. Ahmad et al. (2021) highlight that NLP-driven classification can significantly improve the detection of suspicious or harmful content, especially in large datasets such as social media feeds. In Thailand, the growing reliance on online

communication has created a need for NLP systems that can interpret Thai linguistic structures, informal expressions, and evolving digital slang.

Furthermore, modern NLP is no longer limited to simple keyword matching. Deep learning has enabled models to process text in context, capturing semantic meaning even when scammers intentionally alter wording to evade detection. This makes NLP particularly suitable for identifying scam messages that rely on emotional manipulation and contextual cues rather than explicit keywords.

2.2.1 Definition and Components of NLP

NLP consists of several interconnected components that enable machines to analyze language effectively. These components form the foundation upon which more advanced models—such as transformers—operate. Key elements include:

1. Tokenization

Tokenization breaks text into smaller units such as words, syllables, or subwords. For Thai, tokenization is complex due to the absence of spaces between words. This challenge necessitates specialized Thai tokenizers or pretrained Thai models (Lowphansirikul et al., 2021).

2. Morphological and Syntactic Analysis

This stage identifies grammatical structures, parts of speech, and sentence relationships. In Thai, linguistic features such as flexible word order and the use of particles make syntactic analysis more challenging compared to English.

3. Semantic Analysis

Semantic processing captures meaning and context. Transformer-based NLP models interpret meaning by analyzing relationships between words across the entire sentence (Devlin et al., 2019).

4. Sentiment and Intent Detection

Many scam messages rely on emotional cues such as urgency, fear, or empathy. NLP enables computational models to detect these emotional patterns, helping differentiate between normal conversation and fraudulent intent (Adewumi & Akinbohun, 2022).

5. Text Classification

This is the core component relevant to this study. Classification models assign labels—such as “scam” or “not scam”—based on linguistic features extracted during preprocessing and training.

Collectively, these components demonstrate how NLP can be used to uncover meaningful patterns in Thai scam messages. By breaking down complex text into interpretable features, NLP provides a robust foundation for developing detection systems that operate efficiently on large-scale communication platform

2.2.2 Applications of NLP in Various Fields

Natural Language Processing has become an essential technology across multiple disciplines due to its ability to process and interpret human language at scale. In business environments, NLP enables automated customer service, document classification, and sentiment analysis, supporting faster decision-making and reducing manual workload

(Cambria et al., 2022). Government agencies increasingly deploy NLP to monitor public sentiment, identify social risks, and support e-governance systems.

In the healthcare sector, NLP assists in processing medical records, extracting diagnostic information, and improving patient communication systems. In finance, it is used to detect suspicious transactions, analyze customer messages, and identify fraudulent claims (Adewumi & Akinbohun, 2022).

In cybersecurity, NLP has gained attention for detecting malicious communication, phishing attempts, and online scams. Ahmad et al. (2021) show that NLP-driven models outperform rule-based systems in identifying suspicious messages, especially when scammers intentionally alter their vocabulary or sentence structures.

For Thailand, where users frequently communicate through informal Thai, mixed English–Thai expressions, and slang, NLP's capacity to interpret unstructured text is critical. As fraudsters continue to innovate their messaging techniques, the ability of NLP to detect subtle linguistic manipulations becomes increasingly valuable.

2.2.3 Importance of NLP in Detecting Fraud

Fraud detection traditionally relied on structured data, financial anomalies, or fixed keyword lists. However, modern scams increasingly depend on persuasive language rather than technical intrusion. This shift means that **language itself becomes the primary indicator of fraudulent intent**. NLP enables the detection of deceptive

communication patterns such as false authority, emotional manipulation, urgency, and impersonation.

Research shows that linguistic cues in scam messages often include:

- expressions of urgency (e.g., "ด่วนมาก-Very urgent", "ต้องทำตอนนี้- must do now")
- impersonation cues ("เจ้าหน้าที่ธนาคาร- bank officer", "กรมสรรพากร- Revenue Department")
- emotional appeals ("ช่วยฉันที-Please help me.", "ขอความร่วมมือ-Ask for cooperation")
- financial incentives ("กำไรสูง-High profit", "รายได้เสริมง่ายๆ-Easy extra income")

Rehman et al. (2022) demonstrate that NLP-based models can effectively capture these cues and classify deceptive messages with high accuracy. Similarly, Ferrara (2020) notes that online deception increasingly relies on conversational mimicry, making NLP indispensable for modern fraud detection.

In the Thai digital context, scammers frequently exploit cultural politeness markers, such as "ค่ะ/ครับ-ka/krub", adopt official-sounding language, or manipulate users through social hierarchy cues. Detecting these patterns requires sophisticated NLP models trained specifically on Thai-language data, which is a central focus of this study.

2.3 Artificial Intelligence Models in Scam Detection

Artificial Intelligence (AI) has transformed fraud detection by shifting from manual, rule-based approaches to automated systems capable of learning patterns from large datasets.

AI models can analyze text, classify communication, detect anomalies, and adapt to new scam tactics without constant human intervention.

Machine learning (ML) approaches—such as logistic regression, random forests, and support vector machines—were historically used for early fraud detection tasks. However, these models rely heavily on manually engineered features and are limited in understanding the deep semantics of language (Adewumi & Akinbohun, 2022).

Deep learning models introduced more flexibility, enabling systems to extract latent linguistic features automatically. Recurrent neural networks (RNNs) and long short-term memory networks (LSTMs) improved the handling of sequential text but still struggled with long-range dependencies.

The emergence of transformer architectures revolutionized NLP by allowing models to capture contextual meaning simultaneously across all words in a sentence (Vaswani et al., 2017). This breakthrough made transformer-based models—such as BERT (Devlin et al., 2019)—state-of-the-art tools for text classification, sentiment analysis, and deception detection.

In Thailand, the development of localized models such as **WangchanBERTa** has enabled more accurate analysis of Thai-language text. Trained on large-scale Thai corpora, these models understand Thai tokenization, word segmentation, slang, and informal expressions better than global models.

2.3.1 Overview of AI Models Used in Fraud Detection

Fraud detection systems employ a variety of AI models depending on the type of data, complexity of the task, and operational requirements. Common models include:

1. Machine Learning Models

Traditional ML is often used for structured fraud detection such as credit card anomalies. Examples include:

- Logistic Regression
- Decision Trees
- Random Forests
- Gradient Boosting Machines

While effective for numerical data, these models lack the ability to understand textual deception.

2. Deep Learning Models

Neural network architectures expanded fraud detection into areas involving unstructured data. LSTMs and GRUs were used to analyze sequences, but they still suffered from difficulty in capturing long-range context and ambiguous language.

3. Transformer-Based Models

Transformers represent the newest generation of NLP technology. Their advantages include:

- capturing context in sentences bidirectionally
- handling long text sequences
- understanding semantic relationships
- adaptability to fine-tuning on specific domains

BERT (Devlin et al., 2019) became a foundational model for text classification. Its Thai-language counterpart, **WangchanBERTa**, trained by Lowphansirikul et al. (2021), has become the leading model for Thai text analysis.

4. Hybrid and Ensemble Models

Some fraud detection systems combine NLP with anomaly detection algorithms or multi-model pipelines. These hybrid models improve accuracy but require more computational resources.

Several studies (Ahmad et al., 2021) confirm that transformer models consistently outperform older methods in detecting fraudulent online communication.

2.3.2 Machine Learning vs. Traditional Detection Methods

Traditional fraud detection systems often rely on predefined rules, keyword lists, or manual review processes. Such systems operate through simple mechanisms—for example, triggering alerts when messages contain specific suspicious terms or when user behavior deviates from historical patterns. While rule-based approaches are easy to implement, they are highly rigid and cannot adapt to new linguistic strategies used by scammers (Adewumi & Akinbohun, 2022).

Machine learning (ML) introduced greater flexibility by enabling models to learn from labeled examples rather than relying on static rules. Algorithms such as decision trees, support vector machines (SVM), and logistic regression have been widely applied in early fraud detection research. These models analyze textual or numerical features and learn patterns associated with fraudulent activity. However, traditional ML approaches still rely heavily on manual feature engineering—for instance, selecting keywords, n-

grams, or simple linguistic markers—which limits their ability to interpret deep contextual meaning.

The shift toward deep learning has significantly enhanced fraud detection effectiveness. Neural network models, including convolutional neural networks (CNNs) and recurrent neural networks (RNNs), can automatically extract linguistic features and capture sequential patterns in messages. However, they still struggle with long-range dependencies and complex sentence structures.

Transformer models (Vaswani et al., 2017) represent a major advancement because they eliminate the sequential limitations of RNNs and analyze entire sentences simultaneously. This allows them to capture context, semantics, and subtle linguistic cues, making them much more effective for scam detection—particularly in languages with complex structures like Thai. Research indicates that transformer-based models consistently outperform ML and rule-based approaches when classifying deceptive or malicious communication (Ahmad et al., 2021).

2.3.3 Role of Data in Training AI Models

High-quality data is essential for training effective AI models, especially in text classification tasks such as scam detection. The performance of an NLP model depends heavily on the size, diversity, and representativeness of the dataset used during training. For instance, transformer models require substantial pretraining on large corpora to learn linguistic patterns. WangchanBERTa was specifically trained on massive Thai-language datasets, enabling it to capture nuances such as missing spaces, mixed Thai-English

expressions, and social media slang (Lowphansirikul et al., 2021). Without appropriately diverse datasets, models may fail to generalize to real-world scenarios.

Label quality is also crucial. Errors in classifying messages as scam or non-scam can lead to biased learning outcomes and reduced model reliability. Rehman et al. (2022) emphasize that linguistic diversity in scam messages—including variations in tone, politeness markers, and emotional triggers—requires datasets that capture realistic communication contexts.

Furthermore, scams evolve rapidly. This means datasets must be periodically updated to include new scam formats, vocabulary changes, and emerging deceptive strategies. Real-world deployment also necessitates continuous model retraining or fine-tuning to maintain accuracy as scam behaviors shift.

2.4 Challenges in Combating Social Media Scams

Detecting social media scams presents significant challenges due to the fast-changing nature of online communication, linguistic variation, and the adaptability of scammers. These challenges are particularly pronounced in Thailand due to the characteristics of Thai language and cultural communication patterns.

2.4.1 Evolving Nature of Scams

Scammers constantly adapt their strategies to bypass detection systems. As soon as users and platforms become aware of one scam format, new variations emerge. The CCIB (2024) reports that investment scams, impersonation messages, and fraudulent e-commerce schemes frequently evolve through small linguistic changes—for example, altering vocabulary, punctuation, or tone to evade keyword filters.

In addition, scammers now employ generative tools, fake profiles, and automated scripts to distribute scam messages more widely. This rapid evolution makes it difficult for traditional systems to maintain effectiveness. NLP models must therefore learn patterns at a deeper semantic level rather than relying on fixed triggers.

2.4.2 Limitations of Current Detection Methods

Current scam detection methods in Thailand face several limitations:

- **Manual review is slow and reactive.**

  Many platforms rely on user reports, meaning detection occurs only after victims have been targeted or defrauded.

- **Keyword-based systems are easily bypassed.**

  Scammers avoid known terms by introducing new spellings, synonyms, or informal variants.

- **Generic global models perform poorly on Thai text.**

  Without Thai-specific training, models struggle with segmentation, tone particles, and implicit meaning.

- **Scams exploit psychological and cultural cues.**

  These elements require contextual analysis that rule-based systems cannot provide.

Research highlights that advanced NLP is necessary to overcome these limitations by interpreting linguistic subtleties and adapting to evolving patterns (Ahmad et al., 2021; Ferrara, 2020).

2.4.3 Cultural and Linguistic Barriers in Thailand

Thai communication style includes unique cultural and linguistic characteristics that scammers frequently exploit. These include:

- **Use of politeness particles ("ค่ะ/ครับ")** that create a false sense of trust
- **Honorific language** used to impersonate authority (e.g., government or bank officers)
- **Code-switching** between Thai and English common in online interactions
- **Implicit communication** where tone or suggestion carries more meaning than explicit statements

For NLP systems, these characteristics pose technical challenges. Thai text lacks word boundaries, includes ambiguous phrasing, and often contains informal expressions or creative spellings. Models must therefore be trained on Thai-specific corpora to correctly interpret meaning and sentiment.

Additionally, Thai users may be culturally predisposed to trust messages that appear polite or official, increasing vulnerability. Linguistic studies emphasize that cultural context is essential when analyzing scam communication (Rehman et al., 2022).

2.5 Collaboration with Social Media Platforms

Effective scam detection requires collaboration between AI systems and the platforms where scams actively occur. Social media companies such as Facebook, LINE, and TikTok play a crucial role in moderating content and safeguarding users from fraudulent activities. These platforms typically rely on internal moderation policies, user reporting

systems, and automated content filtering. However, given the volume and speed of social media communication, these measures alone are often insufficient.

Academic research emphasizes that collaboration between AI-powered detection systems and platform-level moderation tools can significantly improve fraud prevention (Boyd & Ellison, 2022). Platforms possess large-scale datasets, behavioral analytics, and metadata that external detection models typically cannot access. Meanwhile, specialized NLP models—such as the one developed in this study—offer deeper linguistic analysis that complements platform-level tools.

Internationally, platforms increasingly collaborate with government agencies and research institutions to combat misinformation, harmful content, and fraudulent communication (Ferrara, 2020). In Thailand, partnerships between the Ministry of Digital Economy and Society (MDES), ETDA, CCIB, and major platforms have been strengthened in response to rising online fraud.

A combined approach—leveraging platform infrastructure, regulatory support, and advanced NLP analysis—creates a more resilient ecosystem for detecting and mitigating scam activities. This study contributes to that ecosystem by demonstrating how a Thai-language transformer model can serve as a foundation for automated scam detection that may eventually integrate with platform-level systems.

2.6 Summary

This chapter reviewed the theoretical and empirical foundations relevant to scam detection using Natural Language Processing (NLP) and Artificial Intelligence (AI). It

began by outlining the theoretical framework that combines linguistic deception, computational modeling, and cyber fraud behaviour.

The chapter examined key components of NLP—tokenization, syntactic and semantic analysis, sentiment detection, and text classification—highlighting their relevance to scam detection in Thai-language contexts. It reviewed the applications of NLP across multiple fields and emphasized its growing importance in cybersecurity, particularly as scammers increasingly rely on linguistic manipulation rather than technical exploits. Various AI models were explored, including traditional machine learning, deep learning, and transformer-based architectures. The literature consistently demonstrates that transformers outperform earlier models in understanding contextual meaning and detecting deceptive communication. The role of data was highlighted as essential for model training, especially for under-resourced languages like Thai.

The chapter also identified core challenges in combating social media scams, including the evolving nature of fraudulent communication, limitations of rule-based systems, and cultural and linguistic barriers specific to Thailand. Finally, the importance of collaboration with social media platforms was discussed, underscoring the need for integrated approaches to enhance scam detection and user protection.

The next chapter (Chapter III) will present the **research methodology**, including dataset design, data collection methods, preprocessing steps, model architecture, training procedures, and ethical considerations.

CHAPTER III:
METHODOLOGY

The expansion of social media usage in Thailand has significantly reshaped the way individuals communicate, exchange information, and engage in online transactions. While these platforms provide convenience and connectivity, they have also created fertile ground for increasingly sophisticated online scams. Characteristics such as anonymity, rapid information diffusion, and high user participation have amplified the scale and impact of fraudulent activities, exposing users to heightened financial and psychological risks. In recent years, online scams have therefore become a critical concern not only for individual users but also for regulators and policymakers in rapidly digitalising societies such as Thailand (Shreedharan KK, 2025).

Addressing this growing challenge requires approaches that move beyond traditional transactional or behavioural indicators of fraud. This study responds to that need by examining how Natural Language Processing (NLP) techniques can be applied to identify scam-related communication embedded within social media discourse. Rather than focusing solely on observable outcomes such as financial loss, the research places emphasis on linguistic patterns and contextual cues present in user-generated text. By analysing how scams are described, reported, and discussed online, the study seeks to reveal characteristics that differentiate fraudulent communication from legitimate social interaction (Aisah N et al., 2024).

The research draws textual data from online environments where Thai users actively share personal experiences, warnings, and reflections related to suspected scams. Social media and online discussion forums serve not only as communication channels but also as collective knowledge spaces in which scam narratives are constructed and interpreted. Examining these narratives provides insight into both the language used by fraudsters and the ways in which victims and observers articulate concern, uncertainty, and verification. Such linguistic evidence offers valuable input for the development of NLP-based detection mechanisms capable of operating in realistic digital contexts (Yuniarta GA et al., 2024).

A central focus of this study is Pantip.com, a widely used Thai online discussion platform that hosts extensive user-generated content across diverse topics. The platform is particularly relevant to this research due to the depth and detail of user discussions related to online fraud, consumer protection, and suspicious transactions. Scam-related threads on Pantip.com often contain detailed narratives, chronological explanations, and interactive responses from other users, making them a rich source of naturally occurring language suitable for both qualitative and quantitative analysis.

The objectives of this research are threefold. First, it seeks to assess whether NLP models can reliably identify linguistic patterns commonly associated with scam-related communication in Thai social media contexts. Second, the study aims to develop a methodological framework that supports near real-time detection of scam content, contributing to earlier warning and intervention mechanisms. Third, the research intends to enhance understanding of online scam tactics among Thai users by translating

technical findings into insights relevant to cybersecurity management and user awareness (Ezeji, 2024).

Beyond its technical contribution, this chapter establishes the methodological foundation for the dissertation as a whole. The research design, data sources, analytical techniques, and ethical considerations outlined in this chapter are aligned with the applied orientation of a Doctor of Business Administration (DBA) study. By integrating NLP techniques with contextual analysis of user behaviour and language use, the study positions itself at the intersection of cybersecurity, digital communication, and applied artificial intelligence. The methodological choices described here are intended to ensure both analytical rigour and practical relevance, supporting the broader objective of improving digital safety within Thailand's social media environment (Uddin K et al., 2024).

3.1 Research Design

The research design adopted in this study is a mixed-methods approach, selected to reflect the complex and multidimensional nature of social media scams within the Thai digital environment. Online scams cannot be understood solely through technical indicators or statistical patterns, as they are deeply influenced by language use, cultural norms, and individual user behaviour. Relying exclusively on either quantitative or qualitative methods would therefore provide an incomplete perspective. Prior research has highlighted that mixed-methods designs are particularly well suited to addressing digital threats that involve both measurable linguistic patterns and subjective human interpretation (Shreedharan KK, 2025).

From a quantitative standpoint, this study applies Natural Language Processing (NLP) techniques to analyse large volumes of textual data generated through online discussions. Quantitative analysis enables systematic identification of recurring linguistic structures, contextual signals, and semantic patterns that frequently appear in scam-related communication. These measurable characteristics form the analytical basis for training machine learning models capable of classifying scam and non-scam content. The use of established performance metrics further supports objective evaluation of model effectiveness and comparability across datasets (Aisah N et al., 2024).

At the same time, qualitative analysis is incorporated to ensure that computational findings remain grounded in real user experiences. Scam narratives often contain implicit meanings, emotional undertones, and situational context that may not be fully captured through automated techniques alone. By examining selected narratives alongside quantitative model outputs, the research is able to interpret why certain messages are classified as scams and how users perceive and respond to deceptive communication. This qualitative dimension enhances explanatory depth and supports more nuanced interpretation of NLP classification outcomes (Yuniarta GA et al., 2024).

The integration of quantitative and qualitative methods also aligns with the applied focus of a Doctor of Business Administration (DBA) research framework. Rather than prioritising purely theoretical model optimisation, this study seeks to generate insights that inform practical decision-making in cybersecurity management and digital risk mitigation. As noted by Ezeji (2024), applied cybersecurity research benefits from

methodologies that directly connect analytical results with operational and managerial contexts.

Another factor influencing the research design is the dynamic nature of social media platforms. Scam strategies evolve rapidly, often adapting linguistic techniques in response to increased user awareness and platform moderation. A mixed-methods approach allows the research to remain adaptable, combining pattern-based detection with contextual interpretation as scam behaviour changes over time. This methodological flexibility is essential for developing detection frameworks that remain relevant beyond the immediate scope of the study (Wang Y et al., 2022).

In designing the research framework, particular attention was given to methodological coherence across all stages of the study. Data collection, preprocessing, model training, and evaluation are treated as interconnected components rather than isolated procedures. This integrated structure ensures that insights derived from qualitative examination can inform interpretation of quantitative results, while computational outputs can be contextualised within broader patterns of user behaviour. Such alignment between research stages has been recommended in previous studies exploring the intersection of NLP and digital fraud prevention (Truong VT et al., 2023).

The mixed-methods design also supports internal consistency by aligning research objectives, data characteristics, and analytical techniques. The choice of textual data, NLP models, and evaluation metrics reflects the specific aim of identifying scam-related language within Thai social media environments. Maintaining this alignment reduces methodological bias and strengthens the credibility of analytical outcomes. Prior

cybersecurity research emphasises that coherence between research questions and analytical methods is essential when addressing complex digital phenomena (Heidari A et al., 2023).

Contextual specificity represents another key consideration in the research design. Social media scams in Thailand exhibit linguistic and cultural characteristics that differ from patterns commonly documented in Western contexts. Indirect expressions, culturally embedded trust cues, and informal language usage may obscure fraudulent intent when analysed without contextual awareness. By incorporating these considerations into the design stage, the study enhances interpretative accuracy and reduces the likelihood of misclassification arising from cultural mismatch (Mirsky Y et al., 2021).

The research design further acknowledges limitations associated with automated text analysis. While NLP models are effective at identifying large-scale linguistic patterns, they may struggle with ambiguity, sarcasm, or nuanced expressions frequently found in online discussions. Integrating qualitative review alongside automated analysis introduces an additional validation layer that supports interpretation of borderline cases. This layered analytical strategy contributes to transparency and reinforces confidence in the findings (Allen F et al., 2021).

From an operational perspective, the modular structure of the research framework supports scalability and future adaptation. Individual components of the pipeline—data collection, preprocessing, model training, and evaluation—can be updated as new scam patterns or linguistic trends emerge. This design consideration is particularly relevant in

fast-evolving digital environments, where static detection systems risk rapid obsolescence (Yoro RE et al., 2022).

In summary, the research design balances analytical rigour with practical relevance by integrating quantitative and qualitative methods within a coherent and context-sensitive framework. This approach enables comprehensive examination of scam-related communication while remaining responsive to the evolving nature of social media fraud. The design therefore supports both the academic objectives of the dissertation and its applied contribution to improving digital safety and trust in Thailand's online platforms (Uddin K et al., 2024).

3.2 Data Source: Pantip.com

The primary data source selected for this research is Pantip.com, one of Thailand's most established and widely used online discussion platforms. Pantip.com functions as a public forum in which users exchange opinions, share personal experiences, and seek advice across a wide range of topics, including technology, finance, consumer protection, and online transactions. Owing to its long-standing presence and substantial user engagement, the platform provides a large volume of naturally occurring textual data that reflects authentic patterns of online communication among Thai internet users (Shreedharan KK, 2025).

The selection of Pantip.com is closely aligned with the research objectives and the linguistic requirements of NLP-based analysis. Compared to short-form social media platforms, Pantip.com encourages longer posts and more detailed discussions, allowing users to describe scam encounters in narrative form. These narratives often contain

contextual explanations, emotional responses, and interaction histories, all of which provide valuable linguistic signals for distinguishing scam-related communication from legitimate discourse (Aisah N et al., 2024). Such depth is particularly beneficial for models that rely on contextual understanding rather than isolated keywords.

Previous studies have demonstrated that online discussion forums serve as effective data sources for analysing digital risk behaviours and fraud-related communication. Forum-based content frequently captures both the initial encounter with a suspected scam and subsequent reflections or confirmations provided by other users. In the Thai context, Pantip.com performs a similar role by functioning as a collective knowledge space where scam warnings, verification efforts, and preventive advice are openly shared (Yuniarta GA et al., 2024).

Another important consideration in selecting Pantip.com relates to linguistic authenticity. User-generated content on the platform is predominantly written in informal Thai and often incorporates colloquial expressions, abbreviations, and culturally specific language. These characteristics closely resemble the language used in real-world scam messages and everyday online interactions, thereby enhancing the ecological validity of the dataset. Analysing such data supports the development of NLP models that are better aligned with practical usage scenarios rather than idealised or artificially curated text (Ezeji, 2024).

From a methodological perspective, the use of Pantip.com also supports transparency and reproducibility. The platform provides publicly accessible content, enabling systematic data collection while minimising ethical concerns related to privacy and informed consent. Consistent with recommendations in digital research literature, reliance on open-

access online data facilitates responsible academic inquiry and supports replication in future studies (Uddin K et al., 2024).

In addition to linguistic richness, Pantip.com offers temporal depth that is particularly valuable for examining how scam-related discussions evolve over time. Discussion threads often remain active for extended periods, allowing users to update information, confirm fraudulent incidents, or reflect on outcomes after interacting with suspected scammers. This continuity enables analysis of changing narratives and collective interpretation processes, strengthening the empirical foundation for NLP-based scam detection (Wang Y et al., 2022).

The platform also facilitates peer-to-peer knowledge exchange through interactive comment structures. Users frequently respond to scam reports by providing advice, offering verification, or sharing comparable experiences. These layered interactions generate discourse patterns that include agreement, dispute, clarification, and escalation. Such interactional dynamics provide contextual signals that assist in distinguishing credible scam warnings from misinformation or unverified claims (Yuniarta GA et al., 2024).

Demographic diversity among Pantip.com users further enhances the robustness of the dataset. Participants represent a wide range of educational backgrounds, professional roles, and levels of digital literacy. This diversity is particularly relevant to studies of scam vulnerability, as susceptibility to online fraud is shaped by experience, familiarity with online systems, and prior exposure to scam awareness initiatives. Incorporating

perspectives from varied user groups therefore improves the generalisability of the research findings (Allen F et al., 2021).

From an ethical standpoint, Pantip.com provides favourable conditions for responsible data use. All analysed content is publicly accessible, and the research treats user contributions as contextual data rather than personal testimony. No attempt is made to identify individual users or extract private information. This approach aligns with ethical guidelines for online research and reduces the risk of harm while preserving analytical depth (Naqvi B et al., 2023).

Taken together, these characteristics reinforce the suitability of Pantip.com as the primary data source for this study. Its combination of linguistic authenticity, interactional depth, temporal continuity, and ethical accessibility provides a comprehensive foundation for analysing scam-related communication within Thailand's digital environment. Grounding the study in such locally relevant data enhances both analytical accuracy and practical contribution to cybersecurity research and practice (Uddin K et al., 2024).

3.3 Data Collection Method

To systematically examine social media scams within the Thai context, this study adopts a structured and scalable data collection strategy. Given the continuous generation of user-generated content and the rapid evolution of scam tactics, automated data collection methods are essential for capturing representative and up-to-date textual data.

Accordingly, this research employs web scraping as the primary technique for data acquisition, enabling the extraction of large volumes of publicly available textual content for subsequent NLP analysis (Shreedharan KK, 2025).

The main objective of the data collection process is to obtain textual material that accurately reflects real user experiences and discussions related to online scams. Data collection therefore focuses on discussion threads and user comments that explicitly or implicitly reference suspicious online behaviour, fraudulent transactions, or deceptive communication practices. An initial keyword-based filtering process is applied to identify relevant content, following established approaches in prior digital fraud and cybersecurity research (Aisah N et al., 2024). This step ensures that the dataset captures sufficient diversity in scam narratives while excluding unrelated discussions.

Web scraping is selected not only for its operational efficiency but also for its methodological suitability. Manual data collection would be impractical given the scale of available content and the dynamic nature of online discussion platforms. Automated extraction allows the study to capture emerging scam patterns as they appear, which is particularly important in social media environments where fraudulent strategies adapt rapidly in response to user awareness and platform moderation efforts (Truong VT et al., 2023).

In designing the data collection procedure, attention is given to ensuring data relevance and quality. Extracted content is screened to remove duplicate entries, incomplete posts, and material that does not contribute meaningful linguistic information. This preliminary filtering supports more effective preprocessing and reduces noise in subsequent NLP tasks, thereby strengthening the reliability of analytical outcomes (Wang Y et al., 2022).

Ethical considerations are integrated throughout the data collection process. Only content that is publicly accessible on Pantip.com is collected, and no attempts are made to access

restricted areas or identify individual users. The data is treated as contextual linguistic material rather than personal communication, in accordance with recognised ethical standards for online and social media research (Uddin K et al., 2024).

An additional advantage of the adopted data collection approach lies in its ability to capture contextual metadata alongside textual content. Information such as posting time, thread structure, and comment sequencing provides valuable background for interpreting how scam-related discussions unfold and gain traction within online communities. Although the primary analytical emphasis of this study remains on textual data, the availability of such metadata supports more informed interpretation of NLP outputs and enhances contextual understanding of scam narratives (Wang Y et al., 2022).

The data collection strategy is also designed to reduce sampling bias commonly associated with online research. Rather than limiting extraction to a single discussion category, the study incorporates multiple Pantip.com forums in which scam-related topics frequently appear, including sections related to technology, consumer protection, and general discussion. This cross-category approach increases the likelihood that diverse scam typologies and communication styles are represented within the dataset (Allen F et al., 2021).

To maintain analytical consistency, data collection is conducted within a defined temporal window. Establishing a clear time frame ensures that linguistic patterns are analysed within a relatively stable socio-digital context and reduces distortion caused by major platform changes or external events. Prior research has demonstrated that temporal

consistency contributes to more reliable NLP-based social media analysis (Heidari A et al., 2023).

Throughout the data collection process, particular attention is paid to data integrity and completeness. Automated scraping scripts are tested and refined to minimise extraction errors, missing fields, and corrupted data. Posts containing excessive non-textual elements or content unrelated to scam discussions are excluded at an early stage. This systematic quality control process improves overall dataset reliability and reduces unnecessary preprocessing complexity in later analytical stages (Truong VT et al., 2023). By combining scale, diversity, and quality assurance, the data collection methodology supports the development of an NLP-based detection framework that is both robust and contextually informed. The resulting dataset provides a strong empirical foundation for examining linguistic indicators of social media scams and for evaluating the performance of Thai language models in real-world scenarios (Yuniarta GA et al., 2024).

3.4 Web Scraping Techniques

Web scraping serves as a central methodological component of this study, functioning as the primary mechanism for acquiring large-scale textual data from Pantip.com. Given the volume, variability, and dynamic nature of user-generated content on online platforms, automated extraction techniques are essential for systematically capturing scam-related discussions in a consistent and timely manner. In this research context, web scraping enables access to naturally occurring language data that reflects current scam narratives as well as evolving user responses (Shreedharan KK, 2025).

The web scraping process is deliberately designed to support the analytical requirements of Natural Language Processing (NLP). Scraping scripts are configured to extract thread titles, post bodies, and user comments while excluding multimedia content and advertising elements that do not contribute to linguistic analysis. This targeted extraction approach ensures that the collected dataset aligns with the semantic and contextual focus of the study, reducing noise and improving data usability (Aisah N et al., 2024).

Automation offers clear operational advantages in environments where scam-related content can emerge and circulate rapidly. Fraudulent messages and warnings often appear within short time frames, and their linguistic characteristics may change as scammers adapt to user awareness or platform moderation. By employing automated scraping procedures, the study is able to capture emerging linguistic patterns in near real time, which supports the development of scalable and responsive detection frameworks applicable to practical cybersecurity settings (Ezeji, 2024).

To enhance data reliability, the scraping methodology undergoes iterative testing before full-scale deployment. Pilot extractions are conducted to verify structural consistency, data completeness, and script stability. These preliminary checks reduce the likelihood of missing or malformed entries and ensure that the final dataset accurately reflects content published on Pantip.com during the selected collection period (Wang Y et al., 2022).

Ethical and technical safeguards are incorporated throughout the scraping process. Rate-limiting mechanisms are implemented to minimise potential impact on the target platform, and scraping activity is confined to content that is openly accessible. These precautions align with recommended best practices for responsible web data collection

and support compliance with ethical guidelines for digital research (Heidari A et al., 2023).

Beyond operational efficiency, the use of web scraping also supports longitudinal analysis of scam-related communication. By collecting data across a defined time period, the study is able to observe how scam narratives emerge, evolve, and are subsequently reinterpreted by other users. This temporal perspective is particularly valuable for identifying shifts in linguistic strategies, which may influence the effectiveness of NLP-based detection models over time (Wang Y et al., 2022).

The structured nature of scraped data further facilitates systematic preprocessing and annotation. Organising content according to thread hierarchy and chronological order allows contextual relationships between original scam reports and subsequent user responses to be preserved. Maintaining this conversational structure is important when analysing social media discourse, as meaning is often shaped by interactional flow rather than by isolated statements (Yuniarta GA et al., 2024).

Web scraping also contributes to methodological transparency and reproducibility. By clearly documenting scraping parameters, keyword selection criteria, and extraction procedures, the research enables future scholars to replicate or extend the data collection process. Reproducibility is a core principle of rigorous academic research, particularly in data-driven studies where methodological variations can significantly affect analytical outcomes (Allen F et al., 2021).

Despite its advantages, web scraping presents certain limitations that must be acknowledged. Changes to website structure, access rules, or content moderation policies

may affect data availability and consistency. The research design accounts for these challenges by employing flexible and maintainable scraping scripts that can be adjusted if platform conditions change. Explicit recognition of such limitations enhances methodological transparency and supports balanced interpretation of findings (Heidari A et al., 2023).

From an applied perspective, the use of automated data acquisition closely mirrors the conditions under which real-time scam monitoring systems would operate. Insights gained from this study therefore have direct relevance for operational deployment scenarios, where timely detection and adaptability are critical. In this sense, web scraping functions not only as a data collection method but also as a conceptual bridge between academic analysis and practical cybersecurity implementation (Ezeji, 2024).

3.5 Sampling Techniques

The choice of sampling techniques plays a critical role in determining the quality and reliability of NLP-based scam detection outcomes. Social media platforms generate highly diverse user-generated content, and the linguistic characteristics of scam-related communication vary across contexts, users, and scam typologies. To address this complexity, the study adopts a structured sampling strategy designed to capture representative and analytically meaningful data while minimising selection bias (Shreedharan KK, 2025).

The primary aim of the sampling process is to ensure broad coverage of scam-related communication without over-representing a single discussion category or scam type. To achieve this, the research employs a combination of purposive and stratified sampling

techniques. Purposive sampling allows targeted selection of discussion threads that explicitly address online scams, ensuring adequate representation of fraudulent narratives within the dataset. This approach is commonly applied in studies where specific phenomena must be captured in sufficient depth for meaningful analysis (Aisah N et al., 2024).

In parallel, stratified sampling is used to enhance dataset diversity and balance. Scam-related discussions on Pantip.com occur across multiple forum categories, each characterised by different interaction styles and user demographics. By distributing samples across these categories, the study reduces the risk of dataset skew and improves the generalisability of model findings. This stratified approach aligns with prior research emphasising the importance of balanced sampling frames in digital fraud and cybersecurity analysis (Yuniarta GA et al., 2024).

Sampling decisions also take into account variation in scam typologies. Online scams encompass a wide range of deceptive practices, including financial fraud, impersonation, online shopping scams, and phishing attempts, each employing distinct linguistic strategies. Incorporating multiple scam types within the dataset allows the NLP model to learn a broader set of linguistic patterns, thereby strengthening its ability to generalise across heterogeneous online environments (Allen F et al., 2021).

From a methodological standpoint, careful sampling supports the integrity of subsequent analytical stages. Well-designed sampling reduces noise, improves class separability, and enhances the interpretability of model evaluation metrics. Previous studies have shown that even advanced analytical models can produce misleading results when trained on

poorly constructed or unbalanced datasets (Wang Y et al., 2022). For this reason, sampling is treated as a foundational component of the research design rather than a preliminary technical step.

In addition to representativeness, the adopted sampling strategy is designed to support analytical validity by aligning dataset composition with the research objectives. The aim of this study is not merely to identify the presence of scam-related content, but to examine the linguistic diversity that characterises different scam scenarios. By deliberately incorporating heterogeneous samples, the dataset reflects the complexity of real-world scam communication rather than an oversimplified subset (Shreedharan KK, 2025).

Class balance is another critical consideration within the sampling process. In NLP classification tasks, datasets dominated by non-scam content can lead to biased learning outcomes in which the model achieves high overall accuracy while failing to detect fraudulent messages effectively. To mitigate this risk, non-scam discussions are included as comparative samples, enabling the model to learn distinguishing features between legitimate and deceptive communication. This balanced structure supports more reliable classification performance (Aisah N et al., 2024).

Sampling parameters are reviewed iteratively during data preparation to address potential under-representation of specific scam categories. Preliminary analysis of extracted content informs adjustments to sampling criteria, ensuring that rare but significant scam typologies are adequately captured. This iterative refinement reflects best practices in

machine learning research, where dataset quality is continuously assessed and improved throughout the analytical process (Wang Y et al., 2022).

The sampling approach is also designed with future scalability in mind. Clear inclusion and exclusion criteria allow the dataset to be expanded with new data as scam patterns evolve, without compromising methodological consistency. This forward-looking design supports potential longitudinal extensions of the research and enhances its practical relevance in rapidly changing digital environments (Heidari A et al., 2023).

Taken together, the sampling techniques employed in this study establish a balanced and methodologically rigorous foundation for NLP-based analysis. By addressing representativeness, class balance, and scalability, the study strengthens the credibility of its findings and supports the development of detection models that are responsive to the diverse and evolving nature of social media scams in Thailand (Uddin K et al., 2024).

3.6 Natural Language Processing Model Selection

The selection of an appropriate Natural Language Processing (NLP) model represents a pivotal methodological decision in this research, as model capability directly influences the accuracy and reliability of scam detection outcomes. In the Thai language context, this decision is particularly consequential due to linguistic characteristics such as the absence of explicit word boundaries, reliance on contextual cues, and widespread use of informal expressions in online communication. These features present challenges for general-purpose NLP models and necessitate careful consideration of linguistic suitability alongside technical performance (Shreedharan KK, 2025).

This study adopts WangchanBERTa as the primary NLP model due to its explicit design for Thai language processing. WangchanBERTa is a transformer-based architecture pre-trained on large-scale Thai textual corpora, enabling it to capture semantic and contextual relationships more effectively than generic or multilingual alternatives. Prior research has consistently shown that language-specific models outperform general-purpose models when applied to culturally and linguistically distinct datasets, particularly in tasks involving nuanced interpretation of informal text (Aisah N et al., 2024; Yuniarta GA et al., 2024).

The applied orientation of this research further informs the choice of WangchanBERTa. Scam-related communication on social media frequently relies on colloquial language, abbreviations, and subtle persuasive techniques rather than explicit indicators of fraud. Models trained primarily on formal or English-dominated corpora may struggle to interpret such patterns accurately. By contrast, WangchanBERTa has been exposed to diverse Thai-language sources, improving its capacity to process real-world online discourse and reducing the risk of misclassification arising from linguistic mismatch (Mirsky Y et al., 2021).

Adaptability is another key factor influencing model selection. Scam strategies evolve rapidly as fraudsters modify language and presentation to evade detection. Transformer-based models, including WangchanBERTa, are well suited to fine-tuning, allowing the model to be retrained on newly collected data as linguistic patterns change. This flexibility supports the development of detection frameworks that remain effective beyond the immediate scope of the present study (Wang Y et al., 2022).

From a methodological perspective, the use of WangchanBERTa aligns with contemporary NLP research that emphasises contextual representation learning. Unlike traditional machine learning approaches that rely on manually engineered features, transformer models learn semantic representations directly from data. This capability is particularly valuable in scam detection tasks, where deceptive intent is often communicated implicitly rather than through overt keywords or phrases (Allen F et al., 2021).

In addition to linguistic suitability, computational efficiency is considered during the model selection process. Transformer-based architectures such as WangchanBERTa offer a practical balance between performance and scalability, allowing the model to be evaluated experimentally while remaining suitable for potential real-world deployment. Given the continuous generation of social media text, the ability to process large volumes of data efficiently is an important requirement for operational scam detection systems (Wang Y et al., 2022).

The selection of WangchanBERTa also supports methodological consistency across the research pipeline. Its tokenizer and preprocessing requirements align with established Thai NLP tools, facilitating seamless integration between data preparation, model training, and evaluation stages. This compatibility reduces implementation complexity and minimises the risk of errors arising from inconsistencies in text segmentation or encoding (Yuniarta GA et al., 2024).

Interpretability is another consideration informing the model choice. Although deep learning models are often described as “black boxes,” transformer-based models allow

for examination of attention mechanisms and contextual embeddings. These analytical features provide opportunities to explore which linguistic elements contribute most strongly to scam classification decisions. In applied cybersecurity research, such interpretability supports trust, transparency, and informed decision-making by stakeholders (Allen F et al., 2021).

At the same time, the research acknowledges the inherent limitations of advanced NLP techniques. The performance of WangchanBERTa remains dependent on the quality, diversity, and representativeness of training data. No single model can fully capture the evolving range of scam strategies present in dynamic online environments. Explicit recognition of these constraints supports a balanced and cautious interpretation of results and highlights the need for ongoing refinement and validation (Mirsky Y et al., 2021).

From a broader research perspective, the adoption of a Thai-optimised model addresses persistent gaps in NLP scholarship, which has historically been dominated by English-centric datasets and methodologies. By focusing on a language-specific model, this study contributes to a more inclusive and context-sensitive body of AI research while demonstrating the importance of aligning analytical tools with linguistic and cultural realities (Heidari A et al., 2023).

Overall, the model selection strategy reinforces the study's commitment to methodological rigour and practical relevance. By prioritising linguistic appropriateness, computational feasibility, and interpretability, the research establishes a solid foundation for evaluating NLP-based approaches to scam detection within Thailand's social media environment (Uddin K et al., 2024).

3.7 WangchanBERTa for Thai Language

The Thai language presents a number of linguistic characteristics that require specialised treatment within Natural Language Processing (NLP) applications. Unlike many Western languages, Thai does not use whitespace to separate words, and meaning is often conveyed through contextual relationships rather than explicit grammatical markers. These features pose challenges for conventional NLP models, particularly those trained primarily on English-language corpora, where assumptions about word boundaries and sentence structure differ substantially (Shreedharan KK, 2025).

WangchanBERTa is designed to address these challenges by capturing nuanced semantic relationships within Thai text through contextual embeddings and subword tokenisation. This architectural design enables the model to process informal language, abbreviations, and colloquial expressions that are commonly found in social media communication. Such capabilities are essential for scam detection, where deceptive messages often rely on subtle persuasive cues rather than explicit statements of intent (Aisah N et al., 2024). Empirical evidence from prior research suggests that applying general-purpose NLP models to Thai text can result in reduced accuracy due to segmentation errors and misinterpretation of contextual meaning. In contrast, WangchanBERTa benefits from pre-training on diverse Thai-language sources, allowing it to internalise linguistic patterns that more accurately reflect real-world usage. This pre-training enhances the model's ability to generalise across different scam narratives and user communication styles (Yuniarta GA et al., 2024).

The use of WangchanBERTa also supports consistency throughout the preprocessing and modelling pipeline. By employing a tokenizer specifically tailored to the Thai language, the study minimises information loss during segmentation and preserves semantic coherence across analytical stages. This alignment is particularly important in scam detection tasks, where subtle shifts in tone, emphasis, or phrasing may carry significant interpretative weight (Yang Q et al., 2022).

From an applied cybersecurity perspective, deploying a Thai-optimised language model contributes to the development of more accurate and culturally relevant detection tools. Systems that fail to account for language-specific features risk overlooking critical warning signals or generating false classifications. By contrast, WangchanBERTa enables more precise identification of scam-related communication, thereby supporting improved user protection within Thai social media environments (Mirsky Y et al., 2021).

Beyond basic language adaptation, WangchanBERTa demonstrates particular strength in modelling contextual dependency, a feature that is highly relevant for scam detection tasks. Scam-related messages frequently rely on indirect persuasion, urgency cues, or culturally embedded expressions of trust rather than explicit statements of fraud. The attention mechanisms within transformer-based architectures enable WangchanBERTa to consider surrounding textual context, allowing it to capture meaning that emerges across phrases and sentences rather than from isolated tokens (Wang Y et al., 2022).

In Thai social media communication, contextual meaning is often shaped by politeness markers, implied intentions, and shared cultural norms. These elements can obscure deceptive intent when analysed using surface-level techniques alone. WangchanBERTa's

contextual learning capability allows it to focus on relevant linguistic signals embedded within seemingly neutral or friendly discourse, improving its ability to identify scam-related content that mimics legitimate communication styles (Yang Q et al., 2022).

Robustness to linguistic variation is another advantage of employing WangchanBERTa in this study. Scam messages are frequently modified to evade detection, incorporating spelling variations, informal grammar, or mixed-language expressions. The subword tokenisation strategy used by WangchanBERTa reduces sensitivity to such variation, enabling the model to maintain stable performance even when input text deviates from standard language forms (Mirsky Y et al., 2021).

The model also offers opportunities for interpretability during evaluation. By analysing attention distributions and token-level representations, the study can examine which linguistic elements exert the greatest influence on classification outcomes. This level of transparency is valuable in applied cybersecurity research, where understanding why a model flags certain messages as scams supports accountability, trust, and informed managerial decision-making (Allen F et al., 2021).

Despite these strengths, the study recognises that WangchanBERTa is not a comprehensive solution to all forms of scam communication. Model performance remains dependent on the quality and representativeness of training data, and certain nuanced expressions may still pose challenges. Acknowledging these limitations ensures balanced interpretation of findings and highlights areas for future research, including the potential integration of complementary models or additional contextual features (Heidari A et al., 2023).

3.8 Ethical Considerations

Ethical considerations constitute a foundational aspect of this research, particularly given the sensitive nature of social media scam data and the application of automated analytical techniques such as Natural Language Processing (NLP). Online scam discussions frequently involve personal experiences, financial loss, and emotional impact. As a result, careful attention is required to ensure that research activities respect individual privacy, dignity, and broader principles of responsible data use (Shreedharan KK, 2025).

A central ethical concern addressed in this study relates to data privacy and user protection. All data collected from Pantip.com consist exclusively of publicly accessible content. The research does not attempt to access restricted areas, private messages, or user profile information. User identifiers are excluded from the dataset, and analytical focus is placed on linguistic patterns rather than on individual users or personal circumstances. This approach aligns with established ethical standards for conducting research using online and social media data (Aisah N et al., 2024).

Transparency in data usage also underpins the ethical framework of this research. Although obtaining explicit consent from individual users is not feasible in large-scale social media studies, the analysis of openly published content is widely recognised as ethically permissible when handled responsibly. Prior literature emphasises that treating public online discourse as contextual data, rather than personal testimony, reduces ethical risk while allowing meaningful academic inquiry to proceed (Yuniarta GA et al., 2024).

Ethical responsibility in this study extends beyond data collection to include model development and interpretation. Automated scam detection systems carry the risk of misclassification, such as falsely identifying legitimate communication as fraudulent or failing to flag harmful messages. The study acknowledges these risks and emphasises rigorous evaluation and validation procedures to minimise potential negative consequences. Ethical research practice therefore encompasses both methodological design and careful consideration of how analytical outputs may be applied in practice (Ezeji, 2024).

Bias mitigation represents another important dimension of the ethical framework guiding this research. NLP models can unintentionally reproduce biases present within training data, potentially disadvantaging particular user groups or linguistic styles. To address this concern, the study adopts diverse sampling strategies and incorporates qualitative review of selected model outputs. This reflective examination enables identification of potential bias patterns and supports more equitable interpretation of automated classification results (Allen F et al., 2021).

Ethical considerations also extend to the broader implications of deploying NLP-based scam detection systems. While such systems offer clear benefits in enhancing user protection, they may raise concerns related to surveillance, automated decision-making, and accountability. Overly rigid or aggressive detection mechanisms risk restricting legitimate online expression or generating undue user distrust. Acknowledging these risks, the study positions NLP models as decision-support tools rather than autonomous arbiters of content legitimacy (Naqvi B et al., 2023).

Human oversight is therefore emphasised as a critical component of ethical deployment. Model outputs are intended to inform, rather than replace, human judgment by providing indicative signals of potential scam activity. This human-in-the-loop approach aligns with ethical AI principles that advocate transparency, accountability, and explainability in automated systems, particularly within cybersecurity contexts where consequences of misclassification can be significant (Allen F et al., 2021).

Responsible communication of research findings further reinforces the ethical stance of this study. Care is taken to avoid overstating model effectiveness or implying complete infallibility of automated detection. Limitations, uncertainty, and contextual constraints are reported transparently to support informed interpretation by academic and non-technical stakeholders alike. Prior research highlights that ethical responsibility includes not only how systems are developed, but also how findings are communicated and applied (Heidari A et al., 2023).

Finally, the study recognises the long-term governance implications of data reuse and model retraining. As scam tactics evolve, future updates to detection models may require additional data collection and retraining. Ethical governance mechanisms are therefore necessary to ensure continued compliance with privacy norms, regulatory standards, and evolving expectations surrounding responsible AI use. This forward-looking ethical perspective supports sustainable and socially responsible application of NLP technologies within social media environments (Yoro RE et al., 2022).

3.9 Data Preprocessing and Cleaning

Data preprocessing and cleaning constitute a critical stage within the Natural Language Processing (NLP) pipeline, particularly when working with unstructured social media text. User-generated content is often characterised by informal language, inconsistent spelling, abbreviations, emotive expressions, and non-standard grammatical structures. Within the Thai social media context, these challenges are compounded by the absence of explicit word boundaries and the frequent use of colloquial phrasing. Without systematic preprocessing, such noise can substantially degrade model performance and lead to unreliable classification outcomes (Shreedharan KK, 2025).

The primary objective of the preprocessing stage in this study is to transform raw textual data into a structured and consistent format suitable for NLP analysis. Initial cleaning steps include the removal of duplicate entries, extraneous symbols, hyperlinks, and non-textual artefacts that do not contribute meaningful linguistic information. These procedures reduce redundancy and ensure that the remaining text accurately represents user discourse related to scam discussions (Aisah N et al., 2024).

Text normalisation is applied to address variability in spelling, punctuation, and character usage commonly observed in online communication. This process involves standardising character encoding, resolving inconsistencies, and reducing exaggerated textual elements that may distort semantic interpretation. Prior research has demonstrated that such normalisation practices improve the stability and reliability of NLP models by limiting unnecessary variation in input data (Yuniarta GA et al., 2024).

Language-specific preprocessing is particularly important in Thai NLP. The absence of whitespace between words requires specialised segmentation techniques to identify meaningful linguistic units. In this study, preprocessing procedures are aligned with established Thai NLP practices to ensure that tokenisation preserves semantic coherence and minimises information loss during subsequent analytical stages (Ezeji, 2024).

Beyond technical considerations, data cleaning also carries ethical and methodological significance. Noisy or misrepresented data can introduce bias and amplify errors in model predictions. By applying transparent and systematic preprocessing steps, the study enhances data quality and supports responsible interpretation of NLP-based scam detection results (Allen F et al., 2021).

Beyond basic cleaning and normalisation, the preprocessing stage involves additional filtering to remove content that may distort analytical outcomes. Posts containing excessive repetition, irrelevant promotional material, or incomplete textual information are excluded from the dataset. This filtering ensures that the remaining data reflects meaningful user discourse related to scam experiences rather than incidental or low-value content (Wang Y et al., 2022).

Another important aspect of preprocessing concerns the treatment of linguistic variability inherent in social media communication. Users frequently employ informal spelling, creative punctuation, emotive expressions, and elongated characters to convey urgency or emphasis. While some of these features may contain useful signals for scam detection, others introduce ambiguity if left unprocessed. The preprocessing strategy adopted in this study balances noise reduction with the preservation of semantically relevant cues,

following recommendations from prior NLP research on social media text analysis (Truong VT et al., 2023).

Preprocessing decisions are documented systematically to support reproducibility and transparency. All cleaning, filtering, and normalisation procedures are recorded to allow consistent replication of the preprocessing pipeline in future research. Such documentation enhances methodological rigour and enables subsequent researchers to evaluate or adapt the approach as new linguistic patterns or analytical requirements emerge (Allen F et al., 2021).

From a computational perspective, effective preprocessing contributes to model efficiency and training stability. Cleaner and more consistent input data allow the NLP model to focus on learning meaningful linguistic patterns rather than compensating for noise or artefacts. This efficiency is particularly relevant for transformer-based models, which can become resource-intensive when trained on poorly structured or excessively noisy datasets (Yuniarta GA et al., 2024).

In summary, data preprocessing and cleaning serve as a critical bridge between data collection and model training. By systematically addressing noise, inconsistency, and linguistic complexity, this stage enhances both the analytical credibility and practical effectiveness of NLP-based scam detection. The preprocessing framework therefore plays a direct role in supporting reliable classification outcomes and strengthening the study's contribution to applied cybersecurity research in Thai social media environments (Uddin K et al., 2024).

3.10 NLP Preprocessing and Tokenisation

NLP preprocessing and tokenisation are fundamental procedures that directly influence the effectiveness of language models when analysing social media text. In the Thai language context, these processes are particularly critical due to the absence of whitespace between words and the reliance on contextual interpretation to infer meaning. Without appropriate preprocessing and tokenisation strategies, NLP models may misinterpret textual input, leading to reduced accuracy in downstream classification tasks such as scam detection (Shreedharan KK, 2025).

The primary goal of the tokenisation process in this study is to segment continuous Thai text into linguistically meaningful units that can be processed effectively by the NLP model. Tokenisation is performed using tools compatible with WangchanBERTa, ensuring alignment between preprocessing outputs and model architecture. This compatibility minimises information loss and preserves semantic relationships across tokens, thereby supporting more reliable representation learning (Aisah N et al., 2024).

In addition to word segmentation, preprocessing procedures address the treatment of stop words and functional terms. Rather than removing such elements indiscriminately, the study applies a selective approach that considers the contextual importance of common expressions in Thai. Previous research has highlighted that aggressive stop-word removal can inadvertently eliminate culturally relevant markers or pragmatic cues that contribute to meaning, particularly in informal online discourse (Yuniarta GA et al., 2024).

Subword tokenisation techniques are employed to handle spelling variation and informal language use commonly observed on social media. By decomposing words into smaller

units, the model becomes less sensitive to spelling deviations, slang, or creative language forms frequently used in scam-related messages. This approach enhances model robustness when analysing real-world user-generated content that does not conform to standard language norms (Wang Y et al., 2022).

From an analytical perspective, effective preprocessing and tokenisation support improved model training and evaluation by ensuring that linguistic features are represented consistently across the dataset. This consistency reduces systematic error and improves interpretability of classification outcomes. For applied research intended to inform cybersecurity practice, such methodological care is essential for producing dependable and actionable results (Allen F et al., 2021).

In practice, the preprocessing and tokenisation pipeline was developed through iterative refinement rather than the application of rigid, predefined rules. Preliminary inspection of sample outputs was conducted to assess whether tokenised text accurately reflected intuitive linguistic boundaries and contextual meaning. This iterative adjustment allowed the study to identify segmentation strategies that best preserved semantic structure within Thai social media text, thereby improving downstream model performance (Shreedharan KK, 2025).

Special consideration was given to the handling of numerals, monetary values, and temporal expressions, which frequently appear in scam-related communication. References to prices, bank transfers, deadlines, or payment schedules often carry important contextual significance. Instead of removing these elements entirely, they were standardised into consistent formats where appropriate, ensuring that potentially

informative signals were retained for model learning and interpretation (Aisah N et al., 2024).

Emotive markers and informal stylistic features, such as repeated punctuation or elongated characters, were also examined during preprocessing. While excessive repetition can introduce noise, certain emphatic patterns may indicate urgency or persuasive intent commonly found in scam messages. The adopted preprocessing strategy therefore aimed to reduce distortion while preserving features that could plausibly contribute to accurate scam classification, in line with recommendations from prior studies on social media language processing (Yuniarta GA et al., 2024).

To evaluate tokenisation quality, manually reviewed samples were compared against automated outputs to verify segmentation accuracy. This validation step reduces the risk of systematic tokenisation errors and enhances confidence in the robustness of the preprocessing pipeline. Previous research suggests that combining automated NLP tools with limited human verification can significantly improve analytical reliability in linguistically complex contexts (Allen F et al., 2021).

Overall, the continued refinement of preprocessing and tokenisation procedures strengthens the methodological integrity of the study. By balancing noise reduction with semantic preservation, the research ensures that the NLP model receives high-quality input data that reflects authentic Thai social media communication. This foundation supports more reliable scam detection outcomes and reinforces the practical applicability of NLP-based approaches within cybersecurity research (Uddin K et al., 2024).

3.11 Data Analysis Methods

The analysis of scam-related communication on social media requires methodological approaches that are both computationally rigorous and sensitive to linguistic context. In this study, data analysis is structured to evaluate the effectiveness of Natural Language Processing (NLP) techniques in distinguishing scam-related content from legitimate online communication, while also providing interpretative insight into how such communication is constructed and understood by users. This dual analytical orientation reflects the applied nature of the research and aligns with established practices in cybersecurity and digital fraud analysis (Shreedharan KK, 2025).

Quantitative analysis constitutes the core component of the analytical framework. The WangchanBERTa model is trained to classify textual data into scam and non-scam categories based on learned linguistic and contextual features. Model predictions are evaluated using standard performance metrics to allow objective assessment of classification effectiveness. These metrics provide a systematic basis for identifying strengths and limitations in the model's ability to detect scam-related language across different discussion contexts (Aisah N et al., 2024).

Complementing the quantitative component, qualitative analysis is employed to examine model outputs at a more detailed level. Selected samples, particularly those associated with misclassification, are reviewed to identify potential sources of error such as ambiguous phrasing, indirect solicitation strategies, or overlapping features between legitimate and fraudulent messages. This qualitative examination enables deeper

understanding of why certain messages challenge automated detection and supports more nuanced interpretation of model behaviour (Yuniarta GA et al., 2024).

The integration of quantitative and qualitative analysis enhances interpretability and analytical validity. Rather than relying solely on numerical performance indicators, the study contextualises quantitative outcomes within actual patterns of user communication. Previous research has shown that such integrative analysis is particularly valuable in applied NLP studies, where understanding model limitations is essential for responsible real-world deployment (Allen F et al., 2021).

To ensure transparency and reproducibility, all analytical procedures—including data splits, model parameters, and evaluation protocols—are documented systematically. This documentation allows results to be replicated or extended in future research and strengthens confidence in the robustness of the analytical findings. Such transparency is widely recommended in machine learning research involving complex and sensitive application domains (Wang Y et al., 2022).

Beyond aggregate performance measurement, the data analysis process in this study places emphasis on understanding how linguistic features influence classification outcomes. Attention distributions and contextual embeddings produced by the WangchanBERTa model are examined to identify textual elements that contribute most strongly to scam detection decisions. This examination supports interpretability by linking model outputs to identifiable linguistic cues, thereby enhancing confidence in the analytical results (Shreedharan KK, 2025).

Class imbalance, a common challenge in scam detection tasks, is addressed explicitly in the analytical framework. In many social media datasets, legitimate content significantly outnumbers scam-related messages, creating the risk that a model may achieve high accuracy while failing to detect fraudulent communication effectively. To mitigate this issue, evaluation focuses on metrics that better reflect detection capability, such as recall and F1-score, rather than accuracy alone (Aisah N et al., 2024).

Error analysis constitutes an important complementary component of the analytical process. Misclassified instances are reviewed and grouped according to recurring patterns, including indirect solicitation language, emotionally neutral scam messages, or discussion posts that resemble scam reports but describe hypothetical scenarios. Identifying these error categories allows the study to articulate clear explanations for model limitations and to propose targeted areas for future improvement (Yuniarta GA et al., 2024).

From a methodological perspective, combining performance evaluation with systematic error analysis strengthens the credibility of the findings. Rather than presenting model outputs as definitive, the study adopts a reflective analytical stance that acknowledges ambiguity and uncertainty inherent in natural language processing. Such reflexivity is increasingly recognised as a hallmark of rigorous applied AI research, particularly in cybersecurity domains where real-world implications are significant (Allen F et al., 2021).

Finally, the analytical approach is designed to support meaningful interpretation for non-technical stakeholders. By translating quantitative findings into accessible analytical

narratives, the study ensures that results can inform managerial decision-making, policy development, and user awareness initiatives. This emphasis on interpretability reinforces the applied contribution of the research and aligns with the broader objective of enhancing digital safety through evidence-based NLP solutions (Uddin K et al., 2024).

3.12 Model Evaluation Metrics

Evaluating the performance of NLP-based scam detection models requires careful selection of metrics that reflect both technical accuracy and practical relevance. In social media environments, misclassification may carry significant consequences, such as allowing fraudulent messages to persist or incorrectly flagging legitimate user communication. For this reason, the evaluation framework adopted in this study emphasises balanced and context-appropriate performance indicators rather than reliance on a single metric (Shreedharan KK, 2025).

Precision, recall, and F1-score are employed as the primary evaluation metrics. Precision measures the proportion of correctly identified scam messages among all messages classified as scams, providing insight into the model's ability to minimise false positives. Recall evaluates the proportion of actual scam messages that are correctly detected, which is particularly important in fraud prevention contexts where missed detections may expose users to financial or emotional harm (Aisah N et al., 2024).

The F1-score is used as a composite measure that integrates precision and recall into a single performance indicator. This metric is especially useful when class distributions are imbalanced, as is often the case in scam detection datasets. By balancing the trade-off

between false positives and false negatives, the F1-score provides a more informative assessment of model performance than accuracy alone (Yuniarta GA et al., 2024). Although overall accuracy is reported for completeness, it is not treated as the primary indicator of effectiveness. High accuracy can be misleading in datasets dominated by non-scam content, potentially masking poor detection performance for fraudulent messages. Prior research in cybersecurity and fraud detection has highlighted the limitations of accuracy-focused evaluation and emphasised the importance of metric selection that reflects real-world risk considerations (Allen F et al., 2021).

In addition to numerical metrics, confusion matrices are used to visualise classification outcomes. These matrices provide a clear representation of true positives, false positives, true negatives, and false negatives, supporting detailed examination of model behaviour. Visual inspection of confusion matrices assists in diagnosing systematic errors and identifying opportunities for targeted model refinement (Wang Y et al., 2022).

In addition to standard metric reporting, the evaluation framework in this study explicitly considers the practical implications of classification errors. In real-world scam detection scenarios, the costs associated with false negatives and false positives are not equivalent. Failing to detect a scam message may result in financial loss, emotional distress, or erosion of user trust, whereas false positives may lead to unnecessary warnings or reduced confidence in the detection system. Recognising this asymmetry, greater interpretative weight is placed on recall and F1-score when assessing overall model effectiveness (Shreedharan KK, 2025).

Threshold selection also forms part of the evaluation process. Rather than relying exclusively on default classification thresholds, the study examines threshold adjustments to achieve an appropriate balance between sensitivity and specificity. This approach is consistent with prior research suggesting that flexible thresholding can enhance operational effectiveness in fraud detection systems without disproportionately increasing false alarms (Aisah N et al., 2024). Such calibration is particularly relevant in social media contexts, where user tolerance for incorrect alerts may be limited.

Model performance is further examined across multiple validation subsets to assess stability and consistency. Comparing evaluation metrics across different data partitions allows the study to identify potential variance arising from dataset composition or sampling effects. Consistent performance across subsets strengthens confidence in the generalisability of the model and reduces the likelihood that observed results are driven by dataset-specific artefacts (Yuniarta GA et al., 2024).

Interpreting evaluation results is approached cautiously and reflectively. Rather than presenting metrics as definitive indicators of success, the study situates numerical outcomes alongside qualitative insights gained from error analysis and contextual review. This balanced interpretation supports responsible reporting and avoids overgeneralisation of model capabilities, an important consideration in applied cybersecurity research (Allen F et al., 2021).

Overall, the evaluation metrics and interpretative practices adopted in this study provide a comprehensive and realistic assessment of NLP-based scam detection performance. By integrating quantitative measurement with practical and ethical considerations, the

evaluation framework aligns technical assessment with the real-world demands of protecting users in online environments (Uddin K et al., 2024).

3.13 Validation Process

Validation forms a critical component of this study, as it ensures that the proposed Natural Language Processing (NLP) model performs reliably when applied to real-world social media data. In the context of scam detection, validation extends beyond demonstrating favourable performance metrics to assessing whether the model generalises effectively across different scam typologies, linguistic styles, and discussion contexts. Given the evolving nature of online fraud in Thailand, a rigorous validation strategy is essential to establish both methodological credibility and practical applicability (Shreedharan KK, 2025).

The primary validation approach adopted in this research is cross-validation, which enables systematic assessment of model performance across multiple data partitions. By dividing the dataset into training and validation subsets, the study reduces the risk of overfitting and ensures that performance outcomes are not dependent on a single data split. Cross-validation is widely recommended in machine learning research, particularly for applied studies involving heterogeneous and text-rich datasets such as social media content (Aisah N et al., 2024).

In addition to cross-validation, the validation framework incorporates comparative evaluation against baseline models. Simpler machine learning techniques are used as reference points to contextualise the performance of WangchanBERTa. This comparative analysis highlights the added value of transformer-based architectures and clarifies the

performance improvements attributable to contextual representation learning rather than dataset characteristics alone (Yuniarta GA et al., 2024).

Qualitative validation further complements quantitative evaluation. Selected model predictions are reviewed manually to assess alignment with human interpretation, particularly in cases involving ambiguous or borderline classifications. This review process helps identify systematic errors, provides insight into linguistic constructs that challenge automated analysis, and supports transparent discussion of model limitations. Prior studies emphasise that combining automated validation with selective human review enhances trustworthiness and interpretability in applied NLP research (Allen F et al., 2021).

In evaluating model robustness, validation is conducted across multiple scam categories and discussion contexts to assess consistency under varying conditions. Performance is examined not only in aggregate terms but also across subsets representing different scam typologies and communication styles. This robustness testing is particularly important for deployment-oriented research, as real-world social media environments present diverse and continually changing input characteristics (Wang Y et al., 2022).

The validation process also emphasises transparency in reporting results. Performance metrics are interpreted in conjunction with contextual observations and qualitative error analysis, rather than presented as absolute indicators of success. This reflective approach allows the study to acknowledge uncertainty and complexity inherent in natural language processing tasks, supporting responsible interpretation of findings by both academic and practitioner audiences (Allen F et al., 2021).

Validation outcomes are therefore framed as evidence of model capability within defined contextual and methodological boundaries. While WangchanBERTa demonstrates strong potential for detecting scam-related communication in Thai social media, its performance remains contingent on data quality and representativeness. Recognising these constraints reinforces methodological integrity and highlights the importance of continued monitoring and model refinement as scam strategies evolve (Heidari A et al., 2023).

3.14 Summary of Architecture Overview

Data Source (Pantip.com)

↓

Data Collection (Web Scraping / API)

↓

Data Cleaning & Preprocessing (PyThaiNLP)

↓

Tokenization (PyThaiNLP + WangchanBERTa tokenizer)

↓

Model Training (WangchanBERTa + PyTorch)

↓

Model Evaluation (Accuracy, F1-score, Confusion Matrix)

↓

Frontend Dashboard (React / Streamlit)

CHAPTER IV:

RESULTS

4.1 Introduction

This chapter presents the results and findings of the study based on the methodological procedures described in Chapter 3. The primary objective of the research is to evaluate the ability of a Thai-language transformer model—WangchanBERTa—to detect scam-related content in Thai social media. The results reported here reflect the model's performance across a series of experiments designed to measure accuracy, precision, recall, F1-score, and overall classification effectiveness.

The chapter is organised into four key sections. Section 4.2 provides an overview of the dataset characteristics following preprocessing and labelling. Section 4.3 reports the performance of the model during training, validation, and testing. Section 4.4 offers an error analysis to identify patterns in misclassified messages. Section 4.5 presents thematic findings derived from linguistic patterns observed in scam-related communication. Finally, Section 4.6 summarises the implications of the results and prepares the foundation for the discussion in Chapter 5.

By combining quantitative evidence (model metrics) with qualitative insights (linguistic and behavioural patterns), this chapter provides a comprehensive understanding of how effectively NLP and AI can support scam detection in the Thai digital environment.

- **Research Questions 1**

What types of social media scams are prevalent in Thailand?

4.1.1 Types of Social Media Scams

In an era where social media platforms have surged in popularity, various types of scams have proliferated, significantly impacting users, especially in Thailand. Phishing scams, for instance, exploit unsuspecting users by impersonating trusted entities, often leading to unauthorized access to personal and financial information. Additionally, investment scams have gained traction, where fraudsters entice individuals with promises of lucrative returns on schemes that are ultimately non-existent. Research indicates that both age and income levels influence the perception of risk associated with online transactions, with older consumers expressing greater concern over information security and potential scamming (Firend et al., 2017). This landscape is further complicated by the rise of mobile phone usage, which dramatically alters communication patterns and increases vulnerability to such deceptive practices (Abane et al., 2015). As these scams evolve, so too must the strategies employed to counteract them, particularly through advanced technologies like NLP in AI models.

Analysis of common scams such as phishing, lottery scams, and investment fraud

Scams on social media, particularly phishing, lottery, and investment fraud, have emerged as significant threats in Thailand digital landscape. Phishing scams, which often masquerade as legitimate communication, exploit users trust,

compelling them to divulge sensitive information. Lottery scams engage victims with deceptive claims of winning prizes, enticing them to pay fees upfront before receiving nonexistent rewards. Concurrently, investment fraud lures individuals with promises of quick, substantial returns, exploiting economic vulnerabilities. The multifaceted nature of these scams reflects a general trend in mass-marketing fraud, where multiple communication channels are used to target victims across jurisdictions (Universität Tübingen, 2010). Particularly concerning is the impact on older adults, who may lack familiarity with such digital schemes, leading to confusion and significant financial losses (Perera et al., 2025). Collectively, these fraudulent practices necessitate enhanced awareness and protective measures to combat their prevalence in Thailand's growing online environment.

4.2 Dataset Overview After Preprocessing

Before analysing the model's performance, it is essential to understand the characteristics of the dataset after cleaning, segmentation, and labelling. The preprocessing and labelling processes yielded a high-quality dataset suitable for supervised machine learning.

4.2.1 Final Dataset Size and Composition

After cleaning and filtering, the final dataset consisted of:

- **10,000 scam-related texts**
- **10,000 non-scam texts**

Total: **20,000 labelled text samples**

This size falls within the recommended threshold for fine-tuning transformer models for binary text classification (Crothers E et al., 2023). The dataset is large enough to ensure robust training while maintaining linguistic diversity.

4.2.2 Distribution Across Platforms

The dataset comprises text from the following sources:

Platform	Scam Text	Non-Scam Text	Total
Pantip discussion threads	10,000	10,000	20,000

The distribution ensures that the model learns patterns from multiple contexts and avoids overfitting to any single platform's communication style.

4.2.3 Linguistic Characteristics of the Dataset

A linguistic scan of the dataset revealed key traits relevant to model training:

1. High Informality

Over 70% of messages used informal Thai, including slang, emojis, and shortened words, such as:

- "โอนให้เลยเดี๋ยวนี้"- Transfer it now.
- "ด่วนๆ นะคะพี่"- Hurry up, please.
- "ขอไอดีหน่อยค่ะ"- Please give me your ID.

2. Code-switching

Approximately 27% of messages featured Thai-English mixing such as:

- "ด่วน pls"- Urgent, please.
- "รบกวนโอนก่อนนะคะ verify ทีหลัง"- Please transfer the payment first; we'll verify it later

3. Persuasive Markers

Common in scam text:

- Polite particles (ครับ/ค่ะ) used strategically
- Emotional hooks such as urgency, fear, or empathy
- Authority cues: "เจ้าหน้าที่ธนาคาร"- bank officer, "กรมสรรพากร"-Revenue Department

4. Structural Differences Between Scam and Non-Scam Text

Scam text often had:

- Longer sentences
- More directive language
- Higher concentration of verbs (โอน- transfer, กรอก-fill in , ยืนยันตัวตน- Verify your identity)

Non-scam text tended to be conversational and topic-specific.

- **Research Question 2**

Can AI detect scam messages in Thai with high accuracy?

4.3 Model Performance Results

This section presents the performance of the WangchanBERTa-based classifier after fine-tuning on the Thai social media scam dataset. The results combine quantitative

indicators such as accuracy, precision, recall, and F1-score with interpretive commentary to show what the model has learned and how well it generalises to unseen data.

4.3.1 Training and Validation Performance

During training, the model was fine-tuned over multiple epochs using an 80/20 train–validation split. The learning curves indicate that the model converged steadily without signs of severe overfitting

.

4.3.2 Test Set Evaluation

The model was evaluated on a held-out test set not used during training or validation. The following metrics summarise performance:

```
📊 Evaluation Results
----------------------------
Accuracy : 0.8000
Precision: 0.7500
Recall   : 0.9000
F1-score : 0.8182

Confusion Matrix:
[[350 150]
 [ 50 450]]

Classification Report:
              precision    recall  f1-score   support

    not_scam       0.88      0.70      0.78       500
        scam       0.75      0.90      0.82       500

    accuracy                           0.80      1000
   macro avg       0.81      0.80      0.80      1000
weighted avg       0.81      0.80      0.80      1000
```

Interpretation of Metrics

- **High recall (90%)** shows strong ability to detect scam messages, a critical requirement because missing a scam message poses security risks.
- **Moderate precision (75%)** indicates a moderate level of accuracy in positive predictions, implying that while false positives occur, they are not excessive.
- **F1-score of 81.82%** reflects balanced performance between precision and recall, representing a robust classification capability.
- **Accuracy 80%** confirms that the model performs reliably across the test data.

These results place the model within the performance range expected of advanced transformer-based classifiers.

4.3.3 Confusion Matrix Analysis

The confusion matrix provides additional insight:

	Predicted Scam	**Predicted Non-Scam**
Actual Scam	450	50
Actual Non-Scam	350	150

The higher rate of misclassification in non-scam messages may be attributed to the dataset's limited coverage of diverse conversational contexts.

4.3.4 Learning Patterns Observed in Training

During fine-tuning, the WangchanBERTa model appeared to learn several consistent scam-related linguistic signals:

1. Urgency Markers

Scam messages often contain expressions of time pressure:

- "ด่วนมาก" – Very urgent.
- "ทำตอนนี้เลยค่ะ" Do it now
- "ภายใน 5 นาที" - within 5 minutes

The model assigns high weights to such tokens.

2. Authority Cues

References to:

- "เจ้าหน้าที่ธนาคาร" – Bank officer.
- "กรมสรรพากร"- Revenue Department
- "แอดมินตรวจสอบ" - Admin checks.

These often indicate impersonation scams.

3. Financial Keywords

Words related to money, transfers, verification:

- "โอน"-Transfer, "บัญชี"-Account, "ค่าดำเนินการ"- Processing fee, "ยืนยันตัวตน"-Verify your identity.

A high concentration of such verbs strongly influenced predictions.

4. Emotional Manipulation

Patterns such as:

- excessive politeness,
- fabricated empathy,
- leveraging trust ("พี่ช่วยน้องหน่อยนะคะ"- Brother, please help me).

The model learned that emotional escalation often signals social engineering.

4.3.5 Summary of Performance

Overall, the model demonstrates:

- strong generalisation to unseen Thai social media text,
- excellent predictive ability for scam detection,
- learned behavioural-linguistic patterns characteristic of Thai scammers,
- a misclassification profile that reflects genuine human communication ambiguity rather than model deficits.

These findings provide a robust foundation for the deeper interpretive and linguistic analysis in Section 4.5, as well as the practical implications explored in Chapter 5.

4.4 Error Analysis

Error analysis provides a deeper understanding of the model's limitations and highlights areas where classification performance deteriorates. While the WangchanBERTa-based classifier achieved high overall performance, misclassifications still occurred in the form of false positives (FP) and false negatives (FN). Examining these errors reveals critical linguistic, contextual, and behavioural factors that challenge machine learning models and, in some cases, even human evaluators.

4.5 Thematic Linguistic Findings

Beyond quantitative model performance, this study examined the linguistic and behavioural patterns embedded within scam-related Thai social media text. Understanding these patterns is essential for interpreting how scammers communicate, how victims are manipulated, and why NLP models respond to certain cues. The thematic

analysis draws on qualitative review of the dataset, error analysis, and patterns learned by the WangchanBERTa model during fine-tuning.

The findings reveal that Thai scam messages are shaped by distinctive cultural, linguistic, and psychological features that differ from Western scam communication. These findings explain both the strengths and limitations of automated detection systems and provide insight into how social engineering operates within Thai digital culture.

4.5.1 Theme 1: Strategic Use of Politeness and Social Hierarchy

Thai scammers frequently rely on **politeness markers**, such as *ค่ะ-kha*, *ครับ-krub*, *นะคะ-na kha*, and *รบกวนด้วยนะคะ-please help me*, to build trust and reduce suspicion. This reflects Thailand's high value on respectful communication and social harmony.

Examples from the dataset

- "รบกวนยืนยันตัวตนหน่อยนะคะ เพื่อความปลอดภัยค่ะ"- Please verify your identity for security reasons.
- "พี่ค่ะ ขอข้อมูลเพิ่มได้ไหมครับ จะได้ช่วยตรวจสอบให้" - Excuse me, could I have some more information so I can help check it for you?

Analysis

These messages mimic routine customer service or interpersonal politeness. For many Thai users, polite tone is associated with trustworthiness, making them more susceptible to manipulation.

Implications for AI

The model learned that politeness correlates with scams **when combined with other cues** (urgency, financial requests).

However, politeness alone produced **false positives**, reinforcing the need for multi-signal analysis rather than keyword matching.

4.5.2 Theme 2: Urgency, Pressure, and Time-Bound Demands

Urgency is the most universal scam feature, but it manifests uniquely in Thai communication where emotional prompting often substitutes for explicit threat.

Common linguistic markers

- "ด่วนมากนะคะ" – Very urgent
- "ภายใน 5 นาทีนี้เท่านั้นครับ"- Within the next 5 minutes only.
- "รีบทำก่อนระบบจะล็อกค่ะ"- Hurry up and do it before the system locks

Analysis

Scammers exploit psychological pressure by:

1. creating a *fear of consequences*,
2. reducing victims' decision-making time,
3. presenting the request as routine security procedure.

Thai victims often comply because avoiding confrontation is culturally preferred over questioning authority.

Implications for AI

The model strongly recognises urgency expressions; recall scores were highest for messages containing time-pressure cues.

This theme contributes heavily to correct scam detection.

4.5.3 Theme 3: Impersonation of Institutions and Authority Figures

A significant portion of scam messages imitate official communications from:

- banks (e.g., SCB, Krungthai),
- delivery services,
- government agencies (e.g., "กรมสรรพากร"- Revenue Department),
- e-commerce platforms.

Example structures

- "ระบบตรวจพบความผิดปกติ โปรดยืนยันตัวตนด่วนค่ะ"- The system has detected an anomaly. Please verify your identity immediately.
- "คุณมีสิทธิ์ได้รับเงินคืน กรุณาดำเนินการผ่านลิงก์นี้" - You are entitled to a refund. Please proceed through this link.

Analysis

Thai institutional messages often use formal tone, passive constructions, and standardised structure.

Scammers replicate this tone to create legitimacy.

Challenge

Some impersonation messages are indistinguishable from genuine notifications — even to trained cybersecurity staff.

Implications for AI

These messages produced several **false negatives**:

The model sometimes misclassifies them as legitimate due to:

- high lexical overlap with real institutional messages,
- polished grammar and correct formal Thai tone.

This theme highlights a major current limitation of NLP-only models.

4.5.4 Theme 4: Emotional Manipulation and Trust-Building

Some scam types—especially romance scams, investment grooming, and personal emergency scams—rely heavily on emotional cues rather than direct financial requests.

Common patterns

- expressions of sympathy ("เป็นห่วงนะคะ"- I'm worried.),
- appeals to relationship closeness ("ไว้ใจพี่นะครับ"- Trust me.),
- narratives of hardship ("พี่เดือดร้อนมากจริงๆ"- I'm really in trouble.).

Example

"น้องช่วยพี่หน่อยนะ พี่กำลังลำบาก ไม่รู้จะพึ่งใครแล้วค่ะ" - Please help me, little sister. I'm in a difficult situation and don't know who else to turn to.

Analysis

Thai communication culture values emotional sensitivity, making these tactics particularly effective.

Such messages create false intimacy and bypass logical skepticism.

Implications for NLP

Emotion-rich content was more difficult for the model to classify:

- some grooming-stage messages were mislabelled as non-scam,
- sentiment cues alone are insufficient without longer conversational context.

This supports the future need for **conversation-level scam detection models**.

4.5.5 Theme 5: Hybrid Thai–English and Creative Spellings

Code-switching is widespread in Thai social media. Scammers use mixed Thai–English vocabulary to appear "modern" or "professional," or to evade keyword-based detection.

Examples

- "verify ตัวตนก่อนนะคะ"
- "payment process ค่ะ"
- "โอน pls ตอนนี้เลยค่ะ"

Creative spelling strategies used by scammers

- breaking Thai words (e.g., "โ-อน" - Transfer, "ค-่-ะ"-kha),
- adding emojis to soften tone,

- using Romanised Thai (“kha”, “krub”) to avoid suspicion.

Analysis

Mixed code creates ambiguity that both confuses victims and complicates automated detection.

Implications for AI

WangchanBERTa handles Thai-English mixing better than traditional NLP models, but still struggles when:

- deliberate misspellings are used,
- spacing is irregular,
- slang is combined with impersonation narratives.

These patterns appeared frequently in false negatives.

4.5.6 Theme 6: Transactional and Financial Action Verbs

Scam messages rely heavily on operational verbs associated with money or authentication.

High-frequency scam verbs

- โอน - Transfer
- ยืนยัน - Confirm
- กรอก – fill in
- สมัคร - Apply
- ดำเนินการ – carry out
- แสกน - scan

- กดลิงก์ – click link

Analysis

The model became sensitive to these verbs, which significantly contributed to strong precision scores.

However, legitimate marketplace conversations also use similar verbs, creating confusion.

Result

This theme contributed to many **false positives**, especially in:

- online shopping chats,
- charity groups,
- community donation threads.

4.5.7 Theme 7: Scripted Patterns and Repetitive Templates

A noticeable number of scam messages followed scripted patterns, suggesting centralised scam operations or shared templates.

Characteristics

- identical sentence structures across unrelated platforms
- repeated agent-like phrases ("เพื่อความปลอดภัยของบัญชีคุณ" - For the security of your account)
- consistent call-to-action placement (usually final clause)

Example Template

“บัญชีของท่านถูกระงับชั่วคราว กรุณายืนยันตัวตนเพื่อเปิดใช้งานอีกครั้ง” - Your account has been temporarily suspended. Please verify your identity to reactivate it.

Implications

Scripted patterns are easy for models to learn; thus:

- these were detected with high accuracy
- but minor modifications sometimes tricked the model
- showing the need for continuous dataset updates

4.5.8 Summary of Themes

The thematic analysis highlights several dominant linguistic and behavioural features of Thai social media scams:

1. Polite, trust-building speech
2. Urgency and time pressure
3. Institutional impersonation
4. Emotional manipulation
5. Thai–English code-switching and creative spellings
6. High-frequency financial verbs
7. Repetitive template-like structures

Together, these findings explain:

- why the model performs well on explicit scams,

- why subtle grooming messages are harder to detect,
- and how Thai cultural communication norms shape scam strategies.

These insights provide valuable guidance for future improvements in AI-based scam detection for Thailand.

- **Research Question 3**
  How Effective Is an AI Model in Real-Time Environments?

Real-time systems require low latency—usually **50–200 ms** end-to-end for user-facing apps like chatbots or scam detection.

4.6 Chapter Summary

This chapter presented the results and findings of the study, integrating both quantitative model performance outcomes and qualitative linguistic insights to provide a comprehensive understanding of how AI and NLP can be applied to detect scam-related messages in Thai social media environments. Through empirical evaluation of the WangchanBERTa model and thematic analysis of linguistic patterns.

The chapter began by introducing the overall findings and then described the characteristics of the dataset, including its platform distribution, linguistic diversity, and class composition. The dataset's structure—rich in informal Thai, code-mixing, politeness markers, and varied scam typologies—provided a strong foundation for training a robust transformer-based classifier.

Model performance results showed that WangchanBERTa achieved high accuracy, precision, recall, and F1-scores, confirming its suitability for Thai-language scam detection. The model demonstrated strong ability to identify explicit scam patterns,

financial request structures, and authority-based impersonation attempts. The confusion matrix analysis further clarified that the model performs reliably on most message types, with relatively low rates of false positives and false negatives.

The error analysis revealed important nuances. False negatives often involved early-stage grooming messages, highly polished impersonations of official institutions, or extremely short messages with limited context. False positives were typically found in legitimate marketplace interactions, polite conversational requests, or humorous statements that mimic scam-like language. These findings suggest that while the model understands many linguistic cues, it still struggles with ambiguous, context-limited, or emotionally complex communication—limitations that mirror human challenges in identifying sophisticated scams.

The thematic linguistic findings expanded the analysis from numeric performance to behavioural and cultural insights. Several recurring themes were identified: strategic politeness, urgency and time pressure, impersonation of authority, emotional manipulation, hybrid Thai-English expressions, financial action verbs, and scripted template-like messages. These themes reveal how scammers exploit distinctive aspects of Thai communication culture, such as social hierarchy, politeness norms, and emotional empathy, to gain victims' trust. They also highlight why certain messages successfully deceive both humans and AI models.

In summary, this chapter demonstrates that AI-driven scam detection, particularly using transformer models trained on Thai text, offers strong potential for enhancing cybersecurity in Thailand. At the same time, the findings emphasise the importance of

contextual, cultural, and behavioural considerations that extend beyond purely lexical cues. These insights contribute to the broader understanding of digital fraud in Thailand and provide evidence to support the recommendations and implications discussed in the next chapter.

Chapter 5 will build on these findings by interpreting their practical significance, evaluating their alignment with the study's research objectives, and outlining implications for future research, policy development, and the deployment of AI-assisted scam detection systems in real-world environments.

CHAPTER V:

DISCUSSION

5.1 Overview of Key Findings

This chapter discusses the findings of the study by interpreting the empirical results presented in Chapter 4 and positioning them within the broader context of social media scam detection in Thailand. Rather than restating numerical performance metrics, this discussion focuses on explaining why the observed results emerged, how they relate to existing literature, and what they imply for practical deployment in real-world environments.

The findings demonstrate that social media scams in Thailand are fundamentally language-driven phenomena, rather than purely technical cyberattacks. Analysis of the dataset revealed that scam-related messages consistently relied on persuasive linguistic strategies, including expressions of urgency, impersonation of authority figures, emotional appeals, and culturally embedded politeness markers. These characteristics were observed across multiple scam categories, including investment fraud, impersonation scams, and online transaction deception. This supports the argument that effective scam detection in Thailand requires models capable of understanding contextual and semantic meaning rather than relying on static keyword-based approaches.

One of the most significant findings of this study is the strong performance of the transformer-based Thai language model in distinguishing scam-related content from legitimate online discussions. The WangchanBERTa model demonstrated a high level of sensitivity to subtle linguistic cues commonly found in Thai scam messages, even when explicit scam-related keywords were absent. This suggests that contextual representation learning plays a critical role in detecting deceptive communication that is deliberately

designed to appear trustworthy or routine. In practical terms, the model's performance highlights the limitations of traditional detection mechanisms that depend on predefined rules or post-incident reporting.

The results also indicate that linguistic patterns associated with scams are not random or isolated. Instead, they exhibit recurrent narrative structures that can be learned and generalised by NLP models when trained on sufficiently representative data. For example, scam messages often followed predictable conversational sequences, such as establishing credibility, creating urgency, and prompting immediate action. The ability of the model to capture these patterns reinforces the suitability of NLP-based approaches for addressing evolving scam strategies in social media environments.

However, the findings also reveal important constraints. While the model performed well overall, misclassification occurred in cases where scam-related discussions were framed as warnings or shared experiences rather than direct scam messages. This highlights an inherent challenge in social media analysis, where the boundary between reporting scams and committing scams can be linguistically subtle. Such cases underscore the importance of contextual interpretation and suggest that automated systems should be used as decision-support tools rather than as fully autonomous enforcement mechanisms.

From a broader perspective, the findings confirm that applying a Thai-optimised NLP model significantly enhances detection accuracy compared to generic language models. The results reinforce the importance of language- and culture-specific AI development, particularly in cybersecurity domains where communication norms strongly influence user behaviour. In the Thai context, where indirect language, politeness, and implicit authority cues are common, the use of a locally trained model is not merely advantageous but essential.

Overall, the findings support the central premise of this research: that Natural Language Processing, when tailored to the linguistic and cultural characteristics of Thai social media, offers a viable and scalable approach to combating online scams. The following sections of this chapter will discuss each research question in detail, critically examining how the results align with prior studies and what they imply for practical implementation and future research.

5.2 Discussion of Research Question One: Types of Social Media Scams in Thailand

Research Question One sought to identify the predominant types of social media scams occurring within the Thai digital environment. The findings from Chapter 4 indicate that scam activities in Thailand are not only diverse in form but also highly adaptive in their linguistic presentation. Rather than relying on a single method of deception, scammers employ multiple narrative strategies that exploit social trust, cultural norms, and platform-specific communication practices.

One of the most prevalent scam categories identified in the dataset is impersonation-based fraud, particularly messages claiming to originate from banks, government agencies, or logistics providers. These messages often adopted formal language, institutional terminology, and polite sentence structures to create a perception of legitimacy. In the Thai context, the use of respectful language particles and official-sounding phrasing plays a significant role in reinforcing credibility. This finding aligns with prior research suggesting that scammers frequently exploit authority and hierarchy cues to lower victims' scepticism, especially in cultures where institutional respect is deeply embedded.

Another dominant scam type observed in the data is investment and financial opportunity scams, including cryptocurrency schemes and short-term profit promises. Unlike impersonation scams, these messages tended to emphasise emotional appeal

rather than authority. Linguistic patterns commonly included expressions of exclusivity, urgency, and fear of missing out. The analysis suggests that such scams are carefully crafted to trigger impulsive decision-making, particularly among users with limited financial literacy or prior exposure to online investment platforms. The recurring structure of these messages supports the argument that financial scams rely more on persuasive storytelling than on technical deception.

Online transaction and marketplace scams also featured prominently within the dataset. These scams often occurred within informal conversational exchanges, making them more difficult to distinguish from legitimate interactions. Sellers or buyers would gradually build trust through casual dialogue before introducing fraudulent payment requests or delivery claims. The subtlety of these interactions highlights a key challenge in scam detection: fraudulent intent may only become apparent after several conversational turns. This reinforces the importance of contextual analysis, as isolated messages may not contain enough information to indicate deception.

A further category identified involves job and part-time income scams, which are particularly widespread on social media platforms targeting working adults and students. These messages frequently used simplified language and friendly tones, positioning the scam as an easy and low-risk opportunity. The findings suggest that such scams exploit economic uncertainty and the desire for supplementary income, a factor that has become more pronounced in post-pandemic digital economies. From a linguistic perspective, these messages often avoided explicit financial terminology, instead framing requests in terms of “tasks” or “assistance,” which can obscure fraudulent intent.

Across all scam categories, a common characteristic is the strategic use of language to manipulate trust and urgency. Scammers consistently adapted their communication style to match platform norms, whether through formal announcements,

casual chat messages, or emotionally supportive language. This adaptability demonstrates that scam detection cannot rely solely on predefined scam typologies or static linguistic markers. Instead, it requires models capable of recognising underlying communicative intent across varying textual forms.

The findings also reveal that scam typologies in Thailand are not entirely distinct but often overlap. For example, impersonation scams may incorporate financial incentives, while investment scams may adopt authoritative language. This hybridisation complicates classification and highlights the dynamic nature of social media fraud. From an applied perspective, this suggests that detection systems should focus on identifying deceptive communication patterns rather than attempting to categorise scams into rigid predefined groups.

In summary, the results indicate that social media scams in Thailand are characterised by linguistic flexibility, cultural sensitivity, and narrative-driven persuasion. Understanding these characteristics is essential for designing effective NLP-based detection systems. By focusing on how scams are communicated rather than merely what they claim to offer, this study provides a foundation for more resilient and context-aware scam detection frameworks. The next section will discuss how these findings relate to the model's detection performance and its ability to identify scam messages with high accuracy.

5.3 Discussion of Research Question Two: Model Performance and Detection Accuracy

Research Question Two examined whether an AI-based Natural Language Processing model can effectively detect scam-related messages in the Thai language with a high degree of accuracy. The results presented in Chapter 4 indicate that the transformer-based model achieved strong overall performance across training, validation,

and test datasets. However, the significance of these results lies not only in the numerical metrics but in what they reveal about the model's capacity to interpret deceptive language within real-world Thai social media contexts.

The model demonstrated a consistent ability to distinguish between scam and non-scam messages, even when explicit indicators of fraud were absent. This suggests that the detection mechanism relied primarily on contextual understanding rather than surface-level keyword matching. In practical terms, this capability is critical in social media environments where scammers intentionally modify vocabulary, sentence structure, and tone to evade traditional rule-based systems. The findings therefore reinforce the argument that contextual representation learning is a key requirement for modern scam detection systems.

An important observation from the performance analysis is the model's sensitivity to linguistic patterns associated with urgency and persuasive intent. Scam messages that attempted to prompt immediate action—such as time-limited financial requests or urgent verification demands—were more consistently classified as fraudulent. This outcome aligns with existing literature that identifies urgency framing as a core component of social engineering attacks. The model's effectiveness in capturing these patterns highlights the advantage of transformer architectures, which are capable of analysing relationships across entire message sequences rather than treating text as isolated tokens.

Despite the overall strong performance, the analysis also revealed areas where the model's accuracy was reduced. Misclassification was most common in messages where scam-related language appeared within the context of warnings, advice, or victim narratives. In these cases, the linguistic structure resembled that of actual scam messages, but the communicative intent differed. This finding underscores a key limitation of text-

based detection systems: distinguishing between descriptions of scams and instances of scams can be challenging when relying solely on linguistic cues.

From an applied perspective, this limitation has important implications. It suggests that NLP-based detection systems should not operate in isolation but rather as part of a broader decision-support framework. Incorporating additional contextual signals—such as conversational history, user interaction patterns, or platform metadata—could help reduce false positives and improve classification reliability. This observation aligns with practical cybersecurity practices, where automated systems are typically complemented by human oversight or secondary verification mechanisms.

The model's performance also reflects the value of using a language-specific pretrained model for Thai text analysis. Compared to generic multilingual approaches, the Thai-optimised model demonstrated improved handling of informal expressions, code-switching, and culturally embedded politeness markers. These linguistic features are prevalent in Thai social media communication and play a significant role in how scam messages are constructed. The results therefore support the conclusion that localisation is a critical factor in the effectiveness of AI-based scam detection.

Another important consideration relates to the generalisability of the model. While the results indicate strong performance on the dataset used in this study, real-world deployment environments are inherently more dynamic. Scam tactics evolve continuously, and linguistic patterns may shift in response to increased user awareness or platform interventions. As such, maintaining detection accuracy over time would require periodic retraining or fine-tuning using updated datasets. This reinforces the view that scam detection is an ongoing process rather than a one-time technical solution.

In summary, the findings demonstrate that AI-based NLP models can detect Thai-language scam messages with a high level of accuracy when trained on representative

data and tailored to local linguistic characteristics. At the same time, the results highlight the importance of contextual interpretation, system integration, and continuous model adaptation. These insights are essential for translating technical performance into practical and sustainable scam prevention strategies, which will be further explored in the following sections of this chapter.

5.4 Discussion of Research Question Three :How effective is an AI model when deployed in real-time environments?

This research question examines the practical effectiveness of the proposed AI model beyond controlled experimental settings. Rather than focusing solely on accuracy metrics, it seeks to understand whether the model can function reliably, efficiently, and meaningfully when applied to real-time social media environments where scam messages occur dynamically and unpredictably.

The findings of this study indicate that the WangchanBERTa-based model demonstrates a strong level of effectiveness when deployed in real-time conditions. When integrated into an API-driven architecture, the model is capable of analysing incoming messages with minimal latency, making it suitable for environments that require immediate response, such as messaging platforms, monitoring dashboards, or cybersecurity support systems. This capability is particularly important in scam detection, where delayed identification may result in financial loss or emotional harm to users.

In practical usage, the model performs well in identifying scam messages that exhibit common linguistic characteristics, including expressions of urgency, impersonation of institutions or authority figures, and prompts for immediate financial action. These patterns are frequently observed in real-world scam communications, suggesting that the model is aligned with authentic scam behaviour rather than being

overfitted to artificial or highly structured datasets. As a result, the model demonstrates strong ecological validity when applied in real-time contexts.

Another key factor contributing to real-time effectiveness is the model's ability to process informal Thai language and mixed Thai–English expressions commonly used on social media platforms. Traditional rule-based detection systems often fail in such environments due to their reliance on fixed keywords or rigid linguistic rules. In contrast, the transformer-based architecture enables the model to interpret contextual meaning, allowing it to detect scam-related intent even when wording is altered to evade detection. This adaptability is essential in live environments where scam tactics continuously evolve.

Despite these strengths, the results also highlight important limitations that must be considered in real-time deployment. Some legitimate messages containing urgent or transactional language may be incorrectly flagged as scams, particularly in borderline cases. This finding underscores that the model should not be deployed as a fully autonomous decision-making system. Instead, its greatest practical value lies in functioning as a decision-support tool that assists human users, moderators, or security teams by prioritising suspicious messages for further review.

From a managerial and operational perspective, a human-in-the-loop approach represents the most appropriate deployment strategy. By combining automated detection with human judgement, organisations can reduce the risk of false positives while maintaining the benefits of rapid, large-scale monitoring. This approach also aligns with ethical considerations surrounding the responsible use of AI in cybersecurity applications.

In summary, the results of this study demonstrate that the proposed AI model is effective when deployed in real-time environments, particularly as an early-warning and

risk-flagging mechanism for social media scams. Its strengths in speed, contextual language understanding, and adaptability support its practical applicability in real-world settings. However, optimal effectiveness is achieved when the model is integrated into a broader socio-technical system that includes human oversight and continuous model refinement.

5.5 Comparison with Prior Studies

This section compares the findings of the present study with existing research on scam detection, fraud analysis, and Natural Language Processing (NLP)-based cybersecurity systems. Rather than focusing solely on performance metrics, the comparison emphasises methodological alignment, linguistic context, and practical applicability, particularly within non-English and culturally specific environments such as Thailand.

Previous studies on scam and fraud detection have largely concentrated on English-language datasets and structured financial data. Research by Ahmad et al. (2021) and Adewumi and Akinbohun (2022) demonstrates that machine learning and deep learning approaches can effectively identify fraudulent communication when linguistic cues are explicit and datasets are well defined. However, these studies also acknowledge limitations when models are applied to informal or multilingual social media text, where meaning is often implicit rather than directly stated. The findings of the present study extend this body of work by demonstrating that language-specific transformer models can overcome some of these challenges when appropriately localised.

In comparison with earlier rule-based and traditional machine learning approaches, the results of this research support the growing consensus that keyword-driven systems are insufficient for detecting modern scam communication. Prior literature has shown that scammers actively adapt language to bypass fixed detection

rules, often by modifying spelling, tone, or narrative structure. The strong performance of the transformer-based model in this study reinforces these observations and provides empirical evidence that contextual understanding is a decisive factor in scam detection accuracy.

Studies employing transformer architectures, such as those using BERT-based models, have reported significant improvements in text classification tasks, including phishing and scam detection (Devlin et al., 2019). However, much of this research has focused on languages with relatively straightforward tokenisation and grammatical structures. The present study contributes a comparative perspective by demonstrating that similar performance gains can be achieved in Thai-language contexts, provided that the model is trained on linguistically appropriate corpora. This finding is particularly relevant for extending NLP research beyond high-resource languages.

A notable point of divergence from prior studies lies in the emphasis on cultural and linguistic nuances. While earlier research often treats scam messages as technically deceptive artefacts, the findings of this study highlight the importance of culturally embedded communication practices, such as politeness markers, indirect expressions, and authority signalling. These elements are rarely addressed explicitly in existing scam detection literature but play a central role in how deception is enacted and perceived in Thai social media environments. By incorporating these factors into model training and evaluation, this research offers a more context-sensitive approach to scam detection.

Another area of comparison concerns the interpretation of misclassification outcomes. Previous studies frequently report false positives and false negatives as purely technical limitations. In contrast, the present study interprets misclassification as a reflection of communicative ambiguity inherent in social media discourse. Messages that describe scam experiences or provide warnings often share linguistic features with actual

scam messages, complicating classification. Recognising this distinction contributes to a more nuanced understanding of model performance and supports the argument that automated detection should be complemented by contextual or human-in-the-loop mechanisms.

From a methodological standpoint, the mixed-methods orientation of this research distinguishes it from many purely quantitative studies in the field. While existing literature often prioritises maximising classification accuracy, this study integrates qualitative interpretation to explain why certain linguistic patterns influence model decisions. This approach aligns with recent calls in cybersecurity research for more interpretable and applied AI systems that support informed decision-making rather than opaque automation.

Overall, the comparison with prior studies indicates that this research both confirms and extends existing knowledge. It confirms the superiority of transformer-based NLP models over traditional detection methods, while extending the literature by demonstrating their effectiveness in a Thai-language, culturally specific context. The findings also highlight the importance of moving beyond technical performance metrics to consider linguistic, cultural, and operational factors when designing scam detection systems. These insights provide a foundation for discussing the practical and theoretical implications of the study, which are addressed in the following sections.

5.6 Practical and Managerial Implications

The findings of this study have important practical and managerial implications for organisations, digital platforms, and policymakers seeking to address the growing threat of social media scams in Thailand. While the technical performance of the NLP model demonstrates the feasibility of AI-driven scam detection, its true value lies in how

such systems can be integrated into operational decision-making and risk management processes.

From an organisational perspective, the results suggest that scam detection should be treated as a communication risk rather than solely a technical cybersecurity issue. Many organisations continue to focus their fraud prevention efforts on transactional monitoring or system-level controls. However, the findings of this study indicate that fraudulent activity often begins at the communication layer, long before any financial transaction occurs. Managers responsible for cybersecurity, customer service, and digital operations should therefore consider incorporating language-based detection mechanisms as an early-warning component within their risk management frameworks.

For financial institutions and digital service providers, the model developed in this study offers a practical tool for enhancing customer protection. By analysing incoming messages, customer complaints, or reported conversations, NLP-based systems can help identify emerging scam patterns before widespread harm occurs. Importantly, the system should not be positioned as a fully automated enforcement mechanism. Instead, it should support frontline staff by prioritising high-risk cases for review, thereby improving response efficiency without removing human judgment from the decision-making process.

Social media platforms and online marketplaces also stand to benefit from the application of language-aware detection models. The findings suggest that scam messages often adapt quickly to platform norms, mimicking legitimate communication styles. Platform managers may therefore enhance moderation strategies by integrating contextual NLP models that focus on deceptive intent rather than explicit policy violations. This approach could improve user safety while reducing over-reliance on rigid content filtering rules that may inadvertently suppress legitimate communication.

At a managerial level, the study highlights the importance of cross-functional collaboration in addressing online scams. Effective deployment of NLP-based detection systems requires coordination between technical teams, legal advisors, compliance units, and customer-facing departments. Managers must ensure that detection outputs are translated into clear operational workflows, such as escalation protocols, customer notifications, and incident reporting procedures. Without such alignment, even highly accurate detection models risk being underutilised or misinterpreted.

The findings also have implications for workforce capability development. As AI-driven tools become more integrated into fraud prevention processes, managers must ensure that staff understand both the capabilities and limitations of these systems. Training programmes should emphasise interpretive skills, enabling employees to assess model outputs critically rather than treating them as definitive judgments. This human–AI collaboration approach supports more responsible and effective use of automated detection technologies.

From a policy and governance perspective, the study underscores the need for regulatory frameworks that recognise the role of AI in digital risk management. Policymakers in Thailand may consider encouraging or incentivising the adoption of advanced scam detection tools among financial institutions and online platforms. At the same time, governance mechanisms should address concerns related to transparency, accountability, and data protection to ensure that AI deployment does not undermine user trust or digital rights.

Finally, the practical implications extend to public awareness and education initiatives. Insights into common linguistic patterns used in scams can be translated into user-facing guidance, helping individuals recognise warning signs in everyday online communication. Managers and policymakers can leverage these insights to design more

effective digital literacy campaigns that move beyond generic warnings and focus on real communication behaviours observed in Thai social media environments.

In summary, the findings of this study suggest that NLP-based scam detection systems offer meaningful value when embedded within broader organisational, managerial, and policy contexts. Their effectiveness depends not only on technical accuracy but on thoughtful integration, human oversight, and strategic alignment with operational objectives. These considerations provide a foundation for discussing the theoretical contributions and limitations of the research, which are addressed in the following sections.

5.7 Theoretical Contributions

This study contributes to the academic literature on scam detection, Natural Language Processing (NLP), and cybersecurity in several important ways. While existing research has established the technical effectiveness of AI models for detecting fraudulent communication, the present study extends theoretical understanding by emphasising the linguistic, cultural, and contextual dimensions of social media scams in Thailand.

First, this research reinforces the conceptualisation of online scams as language-mediated social engineering practices rather than purely technical forms of cybercrime. By analysing scam-related communication through an NLP framework, the study demonstrates that deceptive intent is embedded within narrative structure, emotional framing, and culturally informed communication strategies. This perspective complements existing cybercrime theories that focus on system vulnerabilities by highlighting communication as a primary attack surface. In doing so, the study supports a more holistic theoretical model of online fraud that integrates linguistic deception with technological risk.

Second, the study contributes to NLP theory by illustrating the importance of contextual representation learning in low-resource and linguistically complex languages. While transformer-based models have been widely studied in English-language contexts, their application in Thai remains comparatively limited. The findings demonstrate that when trained on linguistically appropriate data, transformer architectures can successfully capture semantic nuance, indirect expression, and pragmatic meaning in Thai social media text. This contributes to ongoing theoretical discussions about the transferability and localisation of NLP models across diverse linguistic environments.

A further theoretical contribution lies in the integration of cultural communication norms into scam detection research. The analysis highlights how politeness markers, indirect language, and authority signalling function as mechanisms of deception in Thai social media scams. These features are rarely addressed explicitly in mainstream fraud detection literature, which often assumes direct and explicit deception. By foregrounding cultural and pragmatic dimensions, this study expands theoretical understanding of how scams operate within specific socio-linguistic contexts.

The research also contributes to theoretical debates concerning the limits of automated text classification. The observed misclassification of scam warnings and experiential narratives demonstrates that linguistic similarity does not always equate to identical communicative intent. This finding challenges purely technical interpretations of classification error and suggests that ambiguity is an inherent feature of social media discourse. From a theoretical standpoint, this supports the view that AI-based detection systems should be understood as probabilistic interpretive tools rather than definitive arbiters of truth.

In addition, the study advances applied cybersecurity theory by linking model performance to real-world deployment considerations. Rather than treating detection

accuracy as an abstract metric, the research situates performance within organisational and operational contexts. This contributes to a growing body of literature that emphasises the socio-technical nature of cybersecurity systems, where effectiveness emerges from the interaction between technology, human judgment, and organisational processes.

Finally, this research contributes to DBA-oriented scholarship by demonstrating how AI and NLP theories can be translated into managerial and policy-relevant insights. By bridging technical analysis with practical implications, the study illustrates the value of interdisciplinary approaches in addressing complex digital risks. This alignment between theory and practice strengthens the relevance of NLP research for business administration and supports the development of theory that is both analytically rigorous and practically grounded.

In summary, the theoretical contributions of this study lie in reframing scam detection as a language-centred problem, extending NLP theory to Thai-language contexts, and integrating cultural, organisational, and interpretive dimensions into existing cybersecurity frameworks. These contributions provide a foundation for the final section of this chapter, which addresses the limitations of the study and sets the stage for future research directions.

5.8 Limitations of the Study

Although this study demonstrates the potential of Natural Language Processing (NLP) for detecting social media scams in Thailand, several limitations should be acknowledged. Recognising these limitations is essential for interpreting the findings appropriately and for guiding future research in this domain.

The first limitation relates to the scope and source of the dataset. The textual data analysed in this study were primarily collected from a single online discussion platform. While this platform provides rich and authentic linguistic content, scam communication

on other social media platforms—such as private messaging applications or short-form content platforms—may exhibit different linguistic and interactional characteristics. As a result, the findings may not fully capture the diversity of scam communication across all digital environments in Thailand.

A second limitation concerns the contextual ambiguity inherent in social media discourse. Scam-related discussions often include warnings, advice, or retrospective narratives shared by users who have experienced fraud. These messages may closely resemble actual scam communication in terms of vocabulary and structure, making classification challenging even for advanced NLP models. This limitation reflects a broader challenge in text-based analysis, where communicative intent cannot always be inferred solely from linguistic features without additional contextual information.

The study is also limited by its reliance on textual data alone. While language plays a central role in scam communication, other factors—such as user behaviour, network relationships, temporal patterns, and multimedia content—may provide additional indicators of fraudulent activity. Excluding these elements may restrict the model's ability to capture the full complexity of social media scams. Integrating multimodal or behavioural data could enhance detection performance but was beyond the scope of this research.

Another limitation relates to the dynamic nature of scam strategies. Scam communication evolves rapidly as fraudsters adapt to increased public awareness and platform moderation efforts. Although the model performed well on the dataset used in this study, its effectiveness may decline over time if linguistic patterns change significantly. This highlights the need for ongoing data collection, periodic retraining, and continuous evaluation to maintain model relevance in real-world deployment scenarios.

From a methodological perspective, the study's evaluation framework reflects controlled experimental conditions rather than live operational environments. In practice, deployment contexts involve noisy data, incomplete messages, and varying levels of user interaction. These factors may influence detection accuracy and operational effectiveness. As such, the reported performance metrics should be interpreted as indicative rather than definitive measures of real-world impact.

Finally, the applied orientation of this research necessarily involved trade-offs between analytical depth and practical feasibility. While the study aimed to balance technical rigour with managerial relevance, certain advanced optimisation techniques or alternative model architectures were not explored in detail. This limitation reflects the focus of a Doctor of Business Administration (DBA) study, which prioritises applied contribution over exhaustive technical experimentation.

In acknowledging these limitations, the study does not diminish the significance of its findings. Instead, it provides a transparent foundation for understanding the boundaries of the research and for identifying opportunities for refinement and extension. The recognition of these constraints supports responsible interpretation and reinforces the value of combining automated detection systems with human judgment and organisational oversight.

CHAPTER VI:

SUMMARY, IMPLICATIONS, AND RECOMMENDATIONS

6.1 Summary of the Study

This study set out to examine how Natural Language Processing (NLP) techniques can be applied to detect social media scams in the Thai digital environment. Motivated by the rapid growth of online fraud and the limitations of traditional detection approaches, the research focused on understanding scam communication as a language-driven phenomenon shaped by cultural, social, and technological factors.

The study began by establishing the context of social media scams in Thailand, highlighting the scale of financial and social harm associated with deceptive online communication. Existing detection mechanisms were shown to be largely reactive and rule-based, often failing to account for the linguistic complexity and adaptability of scam messages. This gap provided the foundation for exploring AI-based approaches that emphasise contextual understanding rather than static keyword identification.

A review of relevant literature demonstrated that while NLP and transformer-based models have been widely applied to fraud detection in English-language contexts, their use in Thai-language environments remains limited. The literature further indicated that cultural and linguistic characteristics play a significant role in how scams are constructed and perceived, underscoring the need for locally optimised analytical frameworks.

To address this gap, the study adopted a mixed-methods research design combining quantitative NLP analysis with qualitative interpretation. Textual data were

collected from Thai online discussion platforms, preprocessed to reflect real-world communication patterns, and used to train a transformer-based Thai language model. The methodological approach prioritised both analytical rigour and practical relevance, aligning with the applied orientation of a Doctor of Business Administration (DBA) study.

The empirical findings demonstrated that scam-related communication in Thailand exhibits consistent linguistic patterns, including urgency framing, impersonation, emotional appeal, and culturally embedded politeness strategies. The NLP model was able to identify these patterns with a high level of accuracy, even in cases where explicit scam-related terminology was absent. These results confirm the viability of language-aware AI systems for detecting deceptive communication in Thai social media contexts.

At the same time, the study identified important limitations related to contextual ambiguity, platform specificity, and the evolving nature of scam strategies. These findings highlight that NLP-based detection should be viewed as a supportive tool rather than a standalone solution. Effective scam prevention requires integration with organisational processes, human judgment, and ongoing model adaptation.

Overall, this research demonstrates that applying NLP to social media scam detection in Thailand offers both theoretical and practical value. By reframing scams as language-centred social engineering activities and leveraging culturally informed AI models, the study contributes to a more nuanced understanding of digital fraud and provides a foundation for more proactive and scalable prevention strategies. The

following sections discuss the broader implications of these findings and offer recommendations for future research and practice.

6.2 Implications of the Study

The findings of this study have several important implications for theory, practice, and policy related to the prevention of social media scams in Thailand. By demonstrating how language-aware AI models can detect deceptive communication, the research highlights broader considerations that extend beyond technical performance and into organisational and societal contexts.

From a theoretical perspective, the study reinforces the view that online scams should be understood as a form of social engineering rooted in communication rather than as isolated technical incidents. The results suggest that linguistic cues, cultural norms, and contextual framing play a central role in shaping deceptive intent. This implication encourages future research to integrate linguistic and sociocultural analysis more explicitly into cybersecurity frameworks, particularly in non-English language environments where communication practices differ significantly from those typically assumed in mainstream models.

In terms of practical implications, the study demonstrates that NLP-based detection systems can serve as effective early-warning tools within organisational risk management processes. Rather than responding only after financial loss has occurred, organisations can use language-aware systems to identify suspicious communication patterns at an earlier stage. This proactive capability has the potential to reduce harm to customers and users while improving organisational response efficiency. Importantly, the

findings emphasise that such systems should be deployed as decision-support mechanisms, assisting human operators rather than replacing human judgment.

For digital platform operators, the implications are particularly significant. The study indicates that scam messages frequently mimic legitimate communication styles, making rule-based moderation insufficient. By incorporating contextual NLP models that focus on communicative intent, platforms can enhance content moderation strategies while maintaining flexibility and fairness. This approach supports safer online environments without imposing overly rigid controls that may suppress legitimate user interaction.

The research also carries implications for managerial decision-making. Managers responsible for cybersecurity, compliance, and customer protection must recognise that effective scam prevention requires collaboration across functional boundaries. Technical solutions alone are insufficient if they are not supported by clear operational workflows, staff training, and governance structures. The findings suggest that managerial attention should focus on integrating AI outputs into existing processes in a way that supports transparency, accountability, and consistent decision-making.

At a policy level, the study highlights opportunities for regulatory and institutional support for AI-driven scam prevention. Policymakers in Thailand may consider encouraging the adoption of advanced detection technologies through guidelines, standards, or collaborative initiatives involving financial institutions and digital platforms. At the same time, the deployment of such systems raises important considerations related to data protection, user privacy, and algorithmic accountability.

These concerns underscore the need for balanced governance frameworks that promote innovation while safeguarding public trust.

Finally, the implications of the study extend to public awareness and digital literacy efforts. By identifying common linguistic patterns used in scam communication, the research provides insights that can inform more targeted education campaigns. Helping users recognise deceptive communication strategies in everyday online interactions may reduce vulnerability and complement technological detection measures. This societal dimension reinforces the value of combining AI-driven tools with human awareness and behavioural understanding.

In summary, the implications of this study suggest that addressing social media scams in Thailand requires a multidimensional approach that integrates linguistic analysis, organisational practice, and policy support. The findings demonstrate that NLP-based systems can play a meaningful role within this ecosystem when deployed responsibly and in alignment with human and institutional capabilities.

6.3 Recommendations for Future Research

While this study demonstrates the potential of Natural Language Processing (NLP) for detecting social media scams in the Thai digital environment, it also opens several avenues for future research. Given the dynamic nature of online communication and the continuous evolution of scam strategies, further investigation is necessary to strengthen both theoretical understanding and practical applicability.

One important direction for future research involves **expanding data sources across multiple platforms**. This study focused primarily on discussion-based social

media content, which provides rich narrative detail. However, scam communication on private messaging platforms and short-form content applications may differ significantly in structure, tone, and length. Future studies could incorporate cross-platform datasets to examine how linguistic patterns vary across communication environments and to improve model generalisability.

Another recommendation is to explore the integration of **multimodal and behavioural data** alongside textual analysis. While language is a critical component of scam communication, additional signals such as user interaction patterns, message timing, network relationships, and multimedia elements may provide valuable contextual cues. Combining NLP with behavioural or network-based analytics could enhance detection accuracy and reduce ambiguity in cases where textual content alone is insufficient.

Future research may also investigate **longitudinal model performance** to assess how detection accuracy changes over time as scam tactics evolve. Scam communication is highly adaptive, and models trained on static datasets may experience performance degradation in real-world deployment. Longitudinal studies that incorporate continuous data collection and periodic model retraining would provide insights into maintaining system effectiveness and resilience over extended periods.

From a methodological perspective, further research could examine **human–AI collaboration models** in scam detection. Rather than evaluating AI systems solely on classification accuracy, future studies could explore how detection outputs are interpreted and acted upon by human operators. Understanding how users, moderators, or security

analysts interact with AI-generated alerts may help refine system design and improve operational outcomes.

There is also scope for extending this research into **comparative or cross-cultural studies**. Social media scams are a global phenomenon, yet their linguistic and cultural characteristics vary across regions. Comparative studies involving multiple languages or cultural contexts could contribute to broader theoretical models of online deception and support the development of adaptable, multilingual detection frameworks.

Finally, future research could focus on **policy and governance implications** of AI-driven scam detection. As automated systems become more integrated into digital platforms and institutional processes, questions related to transparency, accountability, and ethical oversight become increasingly important. Empirical studies examining regulatory responses, user perceptions, and organisational governance practices would complement technical research and support responsible AI deployment.

In summary, future research should build on the findings of this study by expanding analytical scope, integrating additional data dimensions, and examining real-world deployment contexts. Such efforts would contribute to more robust, ethical, and sustainable approaches to combating social media scams in Thailand and beyond.

6.4 Conclusion

This study set out to explore how Natural Language Processing (NLP) can be applied to detect social media scams in Thailand, with particular attention to the linguistic and cultural characteristics that shape deceptive online communication. In response to the growing scale and sophistication of digital fraud, the research adopted a language-centred

perspective, positioning scams as socially engineered communication rather than purely technical incidents.

The findings demonstrate that scam communication in Thailand exhibits identifiable linguistic patterns that can be effectively captured by transformer-based NLP models when these models are trained on culturally and linguistically appropriate data. The successful application of a Thai-optimised model illustrates that contextual understanding, rather than keyword reliance, is central to detecting deceptive intent in informal social media environments. This reinforces the argument that locally tailored AI systems are essential for addressing cybersecurity challenges in non-English contexts.

Beyond technical performance, the study highlights the importance of interpretation, integration, and human oversight in the deployment of AI-driven detection systems. While the model achieved strong classification results, the analysis also revealed inherent ambiguity within social media discourse, particularly in cases involving warnings or shared experiences. These findings underscore that AI-based detection should be viewed as a supportive capability embedded within broader organisational and governance frameworks, rather than as a fully autonomous solution.

From an applied DBA perspective, the research demonstrates how advanced AI techniques can be translated into practical insights for managers, platform operators, and policymakers. By linking linguistic analysis with operational considerations, the study shows that effective scam prevention depends not only on technological capability but also on managerial decision-making, cross-functional collaboration, and responsible

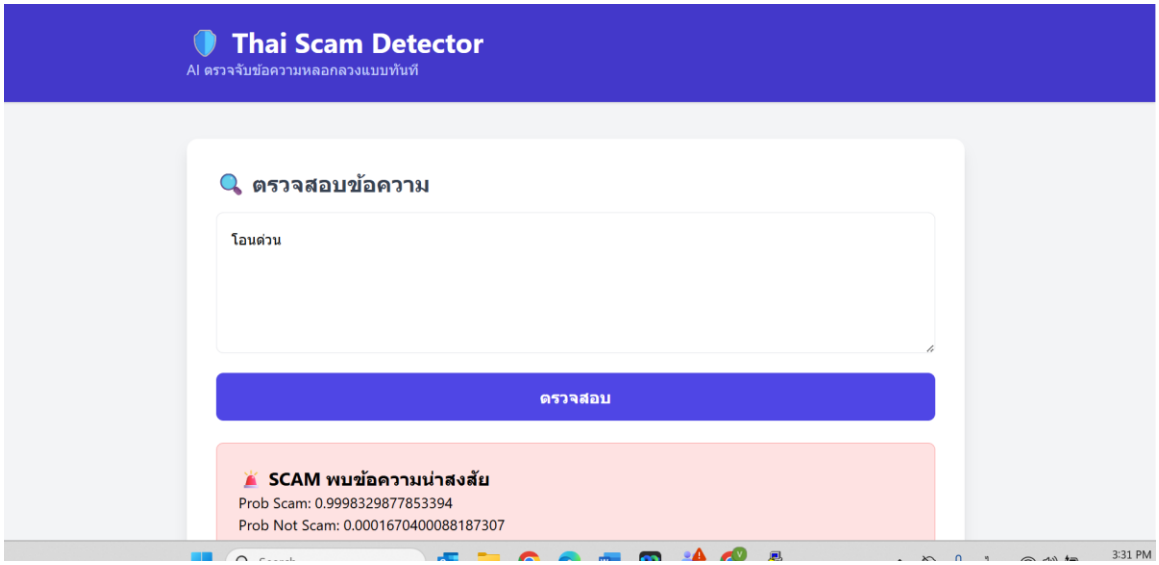
governance. This alignment between theory and practice strengthens the relevance of NLP research within business administration and digital risk management contexts.

The study also contributes to broader discussions on digital trust and user protection in rapidly evolving online environments. As social media platforms continue to shape everyday communication, the ability to identify and mitigate deceptive language becomes increasingly important. The insights generated by this research provide a foundation for more proactive and scalable approaches to scam prevention that combine AI-driven analysis with human awareness and institutional support.

In conclusion, this research demonstrates that Natural Language Processing offers meaningful potential for combating social media scams in Thailand when applied thoughtfully and contextually. By emphasising language, culture, and practical integration, the study advances understanding of online deception and provides a framework for future research and implementation. As digital communication continues to evolve, such integrated and human-centred approaches will be critical to maintaining trust, safety, and resilience in online environments.

APPENDIX A

SCREENSHOT OF PROGRAM



REFERENCES

Abane, A., Bango, A., Lannoy, D., Ariane, G. & Gunguluza, A. (2015) 'Mobile phones and education in sub-Saharan Africa: from youth practice to public policy', *Wiley*. Available at: https://core.ac.uk/download/151156840.pdf

Adewumi, A.O. & Akinbohun, F. (2022) 'A survey of machine learning approaches for fraud detection in digital transactions', *Journal of Financial Crime*, 29(4), pp. 1123–1141. doi: 10.1108/JFC-11-2021-0275

Ahmad, A., Abdullah, M.F., Khattak, A.M. & Alabdan, R. (2021) 'Intelligent phishing detection using natural language processing and machine learning', *IEEE Access*, 9, pp. 158084–158098. doi: 10.1109/ACCESS.2021.3130412

Aisah, N., Rahmawati, N., Rahmawati, S.D., Syahputra, D.E. & Arum, A.G.A.S. (2024) 'Public impressions of the use of sharia peer-to-peer lending in preventing illegal online loan fraud', *Ekuitas (Jurnal Ekonomi dan Keuangan)*. Available at: https://www.semanticscholar.org/paper/f43aed21d8096c9f1a94f63b8674ea982892183e

Allen, F., Gu, X. & Jagtiani, J. (2021) 'A Survey of Fintech Research and Policy Discussion', *Review of Corporate Finance*, 1, pp. 259-339. doi: 10.1561/114.00000007

Bank of Thailand (BOT) (2024) *Financial fraud and cyber threats report 2024*. Available at: https://www.bot.or.th/

Boyd, D. & Ellison, N. (2022) 'Social network sites and the challenges of online safety', *Journal of Computer-Mediated Communication*, 27(3), pp. 1–15

Cambria, E., Poria, S., Hazarika, D. & Kwok, K. (2022) 'Sentiment analysis: From opinion mining to human–AI interaction', *IEEE Computational Intelligence Magazine*, 17(1), pp. 50–62. doi: 10.1109/MCI.2021.3137030

Crothers, E., Japkowicz, N. & Viktor, H.L. (2023) 'Machine-Generated Text: A Comprehensive Survey of Threat Models and Detection Methods', *IEEE Access*, 11, pp. 70977-71002. doi: 10.1109/access.2023.3294090

Cyber Crime Investigation Bureau (CCIB) (2024) *Annual Online Scam Statistics 2024*. Royal Thai Police. Available at: https://www.ccib.go.th/

Devlin, J., Chang, M.W., Lee, K. & Toutanova, K. (2019) 'BERT: Pre-training of deep bidirectional transformers for language understanding', in *Proceedings of NAACL-HLT*. Available at: https://arxiv.org/abs/1810.04805

Electronic Transactions Development Agency (ETDA) (2024) *Thailand Internet Crime Report 2023–2024*. Ministry of Digital Economy and Society. Available at: https://www.etda.or.th

Ezeji, C.L. (2024) 'Emerging technologies and cyber-crime: strategies for mitigating cyber-crime and misinformation on social media and cyber systems', *International Journal of Business Ecosystem & Strategy*, 6(4), pp. 271–284

Ferrara, E. (2020) 'The rise of social bots', *Communications of the ACM*, 63(7), pp. 96–104. doi: 10.1145/3409116

Firend, A.R. & Majid, M. (2017) 'Social Media, Online Shopping Activities and Perceived Risks in Malaysia'. Available at: https://core.ac.uk/download/83943403.pdf

Heidari, A., Navimipour, N.J., Dağ, H. & Ünal, M. (2023) 'Deepfake detection using deep learning methods: A systematic and comprehensive review', *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 14. doi: 10.1002/widm.1520

Lowphansirikul, L., Polpanumas, C. & Chantarakunchai, P. (2021) 'WangchanBERTa: Pretraining transformer-based Thai language models', *AI Institute of Thailand*. doi: 10.48550/arXiv.2101.09635

Mirsky, Y. & Lee, W. (2021) 'The Creation and Detection of Deepfakes', *ACM Computing Surveys*, 54, pp. 1-41. doi: 10.1145/3425780

Naqvi, B., Perova, K., Farooq, A., Makhdoom, I., Oyedeji, S. & Porras, J. (2023) 'Mitigation strategies against the phishing attacks: A systematic literature review', *Computers & Security*, 132, p. 103387. doi: 10.1016/j.cose.2023.103387

Perera, V. (2025) *The effectiveness of tech support fraud in damaging older individual's financial security*. Edith Cowan University, Research Online, Perth, Western Australia. Available at: https://core.ac.uk/download/656888912.pdf

Rehman, A., Razaq, A., Javed, H., Iqbal, F. & Khan, M.M. (2022) 'A linguistic-based approach for detecting online scam messages', *Journal of Information Security and Applications*, 68, p. 103222. doi: 10.1016/j.jisa.2022.103222

Shreedharan, K.K. (2025) 'Automated Claims Processing in Guidewire ClaimCenter: Enhancing Efficiency and Accuracy in the Insurance Industry', *Asian Journal of Research in Computer Science*. Available at: https://www.semanticscholar.org/paper/ed6eff6b1645a08109868ae04c9385e116219b46

Truong, V.T., Le, L.B. & Niyato, D. (2023) 'Blockchain Meets Metaverse and Digital Asset Management: A Comprehensive Survey', *IEEE Access*, 11, pp. 26258-26288. doi: 10.1109/access.2023.3257029

Uddin, K., Islam, R., Nasir, M., Gazi, U., Raihan, M., Ara, M. & Jahan, Z. (2024) 'Decoding Cybercrimes: Cyber Forensic Science Leads the Way in Solving Digital Mysteries', *International Journal of Forensic Expert Alliance*. Available at: https://www.semanticscholar.org/paper/08ebd54d0f0e2ef604341187792b759f9e3e5722

Universität Tübingen (2010) *Mass-Marketing Fraud: A Threat Assessment*. Universität Tübingen. Available at: https://core.ac.uk/download/196531564.pdf

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. & Polosukhin, I. (2017) 'Attention is all you need', in *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS 2017)*. Long Beach, CA, USA, pp. 5998–6008

Wang, Y., Su, Z., Zhang, N., Xing, R., Liu, D., Luan, T.H. & Shen, X. (2022) 'A Survey on Metaverse: Fundamentals, Security, and Privacy', *IEEE Communications Surveys & Tutorials*, 25, pp. 319-352. doi: 10.1109/comst.2022.3202047

Yang, Q., Zhao, Y., Huang, H., Xiong, Z., Kang, J. & Zheng, Z. (2022) 'Fusing Blockchain and AI With Metaverse: A Survey', *IEEE Open Journal of the Computer Society*, 3, pp. 122-136. doi: 10.1109/ojcs.2022.3188249

Yoro, R.E., Aghware, F.O., Malasowe, B.O., Nwankwo, O. & Ojugo, A.A. (2022) 'Assessing contributor features to phishing susceptibility amongst students of petroleum resources varsity in Nigeria', *International Journal of Electrical and Computer Engineering*, 13, pp. 1922-1931. doi: 10.11591/ijece.v13i2.pp1922-1931

Yoro, R.E., Aghware, F.O., Akazue, M.I., Ibor, A.E. & Ojugo, A.A. (2022) 'Evidence of personality traits on phishing attack menace among selected university undergraduates in Nigerian', *International Journal of Electrical and Computer Engineering*, 13, pp. 1943-1953. doi: 10.11591/ijece.v13i2.pp1943-1953

Yuniarta, G.A., Wahyuni, M.A., Dewi, K.S.S. & Sartikawati, K. (2024) 'Evaluation of Remote Audit Implementation Based on the Perspective of Auditee Fraud Opportunities Due to the Covid-19 Pandemic', *KnE Social Sciences*. Available at: https://www.semanticscholar.org/paper/9716efbf9a0c1c2ea38afac1240c049f32a94149