National College of Ireland

Detecting Smishing with NLP, Simulated User Behaviour, and Emotion Analysis: A Machine Learning Approach

MSc Research Project

MSc Cybersecurity

Joshua O'Neill

Student ID: x23315369

School of Computing

National College of Ireland

Supervisor: Dr. Raza Ul Mustafa

National College of Ireland

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Joshua O'Neill
Student ID:	x23315369
Programme:	MSC CYBE_JANO24_O **Year:** 2025
Module:	MSc Research Project
Supervisor:	Dr. Raza Ul Mustafa
Submission Due Date:	11/12/2025
Project Title:	Detecting Smishing with NLP, Simulated User Behaviour, and Emotion Analysis: A Machine Learning Approach

Word Count: 6,657 **Page Count:** 20

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Joshua O'Neill

Date: 10/12/2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Detecting Smishing with NLP, Simulated User Behaviour, and Emotion Analysis: A Machine Learning Approach

Joshua O'Neill x23315369

Abstract

Smishing exploits urgency and emotion to trick users into clicking malicious links or disclosing credentials. Traditional filters that focus on URLs or blocklists struggle with short, rapidly evolving campaigns. This thesis presents an integrated smishing detection framework that combines contextual language modelling, emotion analysis and simulated user behaviour. Four public SMS corpora are merged into a unified dataset, pre-processed with consistent normalisation and evaluated under both balanced and realistically imbalanced splits. Classical TF-IDF baselines with Logistic Regression, Linear SVC and Multinomial Naive Bayes are compared against fine-tuned BERT and DistilBERT models. A late-fusion layer then concatenates GoEmotions probability vectors with BERT or TF-IDF smish scores and trains a Logistic Regression classifier, while a lightweight persona module translates outputs into risk ratings for different user types. Experiments show that BERT with a 256-token context achieves an F1 score of 0.9874 on a deployment-style test set, outperforming TF-IDF baselines. Emotion fusion delivers comparable accuracy while adding interpretable explanations, demonstrating that combining language, emotion and behaviour is a promising direction for human-centred smishing defence.

1. Introduction

Smishing, or SMS based phishing, uses short, urgent and personal messages to deceive recipients into clicking malicious links, disclosing credentials or downloading malware. Recent reports show that smishing is one of the most persistent forms of social engineering across Europe and that the human element, including social engineering and human error, remains the dominant factor in breaches, accounting for around two thirds of incidents (ENISA, 2024; Verizon, 2025). In Ireland, the national postal service has repeatedly issued public warnings about fraudulent customs payment texts that closely mimic legitimate notifications (Scam Alerts – Stay Vigilant | An Post, 2024).

Technical controls have struggled to keep pace with this shift to mobile channels. Blacklists and traditional URL based defences often react too slowly to new campaigns. A study of client-side anti-phishing tools found that around 75% of users could still access phishing sites before blocklisting occurred, and that client side systems detected only about 14% of phishing websites and as few as 10% of login-based attacks in the first hour (Pourmohamad et al., 2024). Attackers further evade detection through URL shorteners, redirects and cloaking that conceal malicious destinations, along with rapid domain churn that outpaces updates. These findings

expose the limits of static filtering and emphasise the need for adaptive detection methods that can interpret linguistic and human centric cues rather than relying only on known bad links.

Smishing is effective because it directly exploits human behaviour. Recent work shows that attackers design messages using cues such as urgency, trust and authority to increase the likelihood of user response, with time sensitive or action-oriented messages generating significantly higher engagement than neutral ones (Timko, Castillo and Rahman, 2023). Concise SMS messages that warn of blocked accounts or delayed deliveries and urge immediate action are reported by recipients as creating a sense of pressure to respond quickly, making them a key cue in smishing attempts (Tabassum, Faklaris and Lipford, 2024). This persuasive effect is amplified by the immediacy of SMS, which is typically read within minutes of delivery and is often perceived as personal and trustworthy (Dobson et al., 2024).

Despite this, most detection systems still treat all users as if they behave in the same way. In practice, susceptibility varies significantly. Timko, Castillo and Rahman (2023) found that participants correctly identified fake SMS messages in 67.1% of cases but genuine ones in only 43.6%, highlighting an asymmetry in human judgement that attackers can exploit. Their study combines content-based machine learning with user experiments and focuses in particular on smishing campaigns that abuse remote assistance tools. While their models achieve strong accuracy on curated datasets and reveal important patterns in how people evaluate suspicious messages, they rely mainly on predefined textual features and specific scam types and do not explicitly incorporate fine grained emotional signals or general user risk profiles. This leaves space for more flexible frameworks that better reflect how different users experience risk.

The motivation for this study comes from this gap between technical detection and human variability. Existing approaches often focus narrowly on content features, achieving strong benchmark accuracy but failing to address persuasive tactics or human variability. This project seeks to bridge that divide by combining contextual Natural Language Processing, emotion analysis and simulated user behaviour into a unified framework. The system applies Natural Language Processing to capture contextual meaning, emotion analysis to identify persuasive cues and simulated user behaviour to reflect differences in risk perception. By aligning technical accuracy with human interpretability, the study aims to deliver a defence that is both effective and sensitive to how different users experience risk in real world settings.

Within this aim, the thesis makes three main contributions while also acknowledging key limitations. It develops an end to end smishing detection pipeline that combines contextual language features, emotion probabilities and persona based interpretation; evaluates classical baselines, transformer models and fusion models on a unified corpus that merges several public SMS datasets under both balanced and realistically imbalanced conditions; and introduces a persona layer that translates model outputs into risk scores and explanations for different user types. At the same time, the framework is constrained by its reliance on public SMS corpora, emotion features derived from a model trained on social media rather than SMS and persona definitions based on expert judgement rather than new user studies, so it should be viewed as an exploratory framework rather than a production deployment.

1.1 Research Statement

This study addresses the challenge of detecting smishing messages that exploit emotional manipulation and user behaviour beyond simple text analysis. It aims to develop an integrated framework that combines contextual language understanding, emotion analysis, and simulated

behavioural profiles to improve detection accuracy and interpretability. The system will be evaluated against classical and transformer-based models using metrics such as precision, recall, F1 score, accuracy, and AUC. This research contributes a practical and explainable defence framework that aligns with the objectives of model benchmarking, system design, and performance evaluation.

1.2 Research Objectives

1. Review existing models and methods for smishing detection, identifying their limitations in handling message context, emotional manipulation, and user diversity. This will include a critical analysis of feature based, machine learning, and transformer-based approaches, with attention to reported evaluation metrics such as accuracy, precision, recall, and F1.

2. Design and implement an integrated detection framework that combines Natural Language Processing for contextual understanding, emotion analysis to capture persuasive cues, and simulated user behaviour to reflect diverse patterns of susceptibility.

3. Develop the system architecture and components to support this framework, ensuring scalability to large datasets, adaptability to evolving smishing tactics, and interpretability so results can be clearly understood by both researchers and practitioners.

4. Evaluate and benchmark the proposed system against classical machine learning baselines and transformer-based models, using standard performance metrics (accuracy, precision, recall, F1 score, and AUC). Additional experiments will test each component individually to show how much emotion analysis and user behaviour modelling improve results, addressing gaps that remain in existing approaches.

Together, the research statement and objectives translate the high-level problem of smishing into a concrete research plan. Chapter two surveys relevant work on smishing detection using NLP and transformer models, emotion recognition and user behaviour modelling, and concludes with the research gap that motivates the unified framework developed in this thesis.

2. Literature Review

This chapter reviews academic and industry work that underpins the proposed smishing detection framework and is organised into three related subsections. Section 2.1 traces the evolution of NLP approaches to smishing and SMS spam, from lexical methods to transformerbased models, focusing on how they represent messages, build datasets, and report performance**.** Section 2.2 reviews research on emotion recognition and user behaviour in adversarial or persuasive text, including phishing and other social engineering attacks. Section 2.3 then draws together the main limitations across these strands and explains how this thesis moves beyond the current state of the art. The scope is deliberately confined to these three strands because they correspond directly to the three layers of the artefact developed in this project.

2.1 Evolution of NLP Methods for Smishing Detection

Smishing detection is uniquely challenging because SMS messages are short, informal and often emotionally manipulative. One recent study evaluated detection using the SMS Spam Collection dataset of 5,574 labelled messages and applied a Random Forest classifier with TFIDF features, achieving 98 % accuracy but only 86 % recall for spam, meaning nearly one in five malicious texts were missed (Saputra, 2024). The contribution of this work lies in demonstrating the speed and accessibility of lexical methods while revealing their weakness: reliance on surface word patterns that fail against obfuscation, synonym variation and persuasive phrasing. For this project, that finding establishes a rationale for moving beyond static feature-based techniques toward contextual and semantic understanding.

Building on this, a hybrid CNN–BiGRU architecture has been proposed to capture both spatial and temporal dependencies in SMS data. Trained on four combined datasets, the model achieved up to 99.82 % accuracy and used LIME visualisation to explain predictions (Mahmud et al., 2024). The contribution of this study is twofold: it improved accuracy through deep contextual learning and introduced interpretability into smishing detection. However, the model was still evaluated only on static, pre collected datasets, limiting its real-world adaptability. This highlights the importance of cross-dataset evaluation and transparency, both of which are central aims of this thesis.

Another strand of work integrates structural information into transformer models. One example combines transformer embeddings with graph convolutional networks to represent semantic and syntactic relations, reaching 99.28 % accuracy on the UCI_SMS dataset but dropping to 93.78 % on the more balanced ExAIS_SMS corpus (Shen et al., 2025). The main contribution here is to show that structural context improves detection beyond pure semantics, while the drop in performance across datasets indicates sensitivity to distributional shifts. For the present research, this underscores the need to benchmark across diverse datasets and to design models that remain robust when message distributions change.

Transformer based models have since become dominant in smishing and phishing research more broadly. One study combined semantic tagging, stylistic cues and contextual embeddings to achieve 97.89 % accuracy and an F1 score of 0.963 across five public SMS datasets, demonstrating that transformers can capture subtle cues such as authority, urgency and curiosity that are central to social engineering attacks (Hosseinpour and Das, 2025). Another study fine-tuned DistilBERT and RoBERTa for phishing detection and used Adaptive Synthetic Sampling (ADASYN) to balance classes, achieving 99 % accuracy on balanced datasets (Jamal, Wimmer and Sarker, 2024). Together, these works show that transformers offer strong benchmark performance and rich linguistic representations, but they also highlight dependence on well curated and often artificially balanced data. This project therefore builds on their insights by testing transformer models under more realistic, imbalanced conditions.

Multilingual and low resource contexts have been explored less thoroughly, despite their practical importance. Research that combined BERT with a character level CNN to detect Bangla smishing achieved 98.47 % accuracy, demonstrating that transformer frameworks can adapt effectively to non-English, low resource languages (Tanbhir *et al.*, 2024). However, performance in that study relied on a relatively small, manually curated corpus. For this thesis, those findings mainly motivate the need for larger and more broad datasets, even within English, so that reported results are less dependent on a single narrow benchmark.

Dataset quality itself significantly shapes reported performance. The SMS Spam Collection remains the most widely used benchmark dataset, but its 5,574 messages are now outdated and English only (Almeida, Hidalgo and Yamakami, 2011). Newer corpora such as SmishTank and SpamDam expand the data pool with 1,062 and 76,577 messages respectively, improving recency and scale (Timko and Rahman, 2024; Li et al., 2024). Yet both are limited because they contain only smishing and no ham, preventing direct evaluation under realistic class imbalance. The present research addresses these issues by merging multiple public datasets, cleaning and normalising them into a unified corpus, and evaluating models under both balanced and imbalanced splits that better reflect operational conditions.

In summary, work on NLP for smishing has evolved from lexical feature extraction to transformer-based architectures that capture contextual, structural and, to some extent, emotional nuance. Across these studies, contributions include improved accuracy, interpretability and multilingual adaptability. However, common limitations remain in generalisation across datasets, handling real world class imbalance and reliance on static, narrow benchmarks. These challenges motivate a methodology that integrates linguistic, emotional and behavioural layers to strengthen interpretability and resilience in practical smishing detection.

2.2 Emotion and Behavioural Modelling in Adversarial Text

Emotion detection provides a psychological dimension to textual analysis that is particularly relevant to social engineering. One comparative study examined lexicon-based sentiment detection alongside classical machine learning models such as Naive Bayes and SVM, finding that both performed noticeably worse than a deep learning architecture that combined convolutional and recurrent layers (Machová et al., 2023). The hybrid one dimensional Convolutional Neural Network Plus Long Short Term Memory model achieved an overall accuracy of 91 %, capturing contextual dependencies and subtle affective cues. This evidence supports the case for neural architectures that can recognise nuanced or ambiguous emotions in short conversational text and, by extension, in SMS smishing messages.

Dimensional emotion frameworks such as Valence–Arousal–Dominance (VAD) provide theoretical grounding for understanding the emotional triggers exploited in smishing. The VAD model represents emotions in a multidimensional space of valence, arousal and dominance, enabling nuanced measurement of affective tone in text (El-Haj and Takanami, 2023). For example, a fraudulent tax refund message typically exhibits high arousal and negative valence, signalling urgency and fear. The structured representation of affect provided by VAD supports the integration of emotion probabilities into detection systems, enhancing their ability to capture manipulative or deceptive tone.

A large-scale systematic review of 226 emotion recognition studies published between 2013 and 2023 found that deep learning models consistently outperformed traditional machine learning approaches across benchmark datasets (Chutia and Baruah, 2024). That review identified Recurrent Neural Networks, Gated Recurrent Units and transformer based models as the dominant methods for emotion analysis, while highlighting ongoing challenges around adversarial manipulation, data dependency and evolving language. For this project, those conclusions support using the GoEmotions transformer model to extract probabilistic emotion features from SMS content as part of the detection pipeline.

Recent shared tasks have also demonstrated the potential of multilingual transformer architectures. At the 2025 SemEval competition, adapter based fine tuning of multilingual transformer models such as XLM-RoBERTa and Afro-XLMR achieved macro F1 scores up to 0.84 and strong performance across several low resource African languages (Laureano De Leon et al., 2025). Complementary work on multilingual multi label emotion detection and emotion intensity prediction combined fine-tuned multilingual transformers with instruction tuned large language models, achieving top four rankings across many languages and outperforming XLMRoBERTa baselines in several settings (Xue et al., 2025). These studies are primarily methodological, showing that multilingual transformer architectures can capture nuanced affect across cultures, but they also emphasise significant computational cost and sensitivity to label imbalance. This motivates the use of lighter models such as DistilBERT in the present project for practical deployment.

Only a relatively small body of work has examined the role of emotions directly in cybersecurity contexts. Scenario based email phishing experiments using both common and COVID-19 themed messages have shown that fear, anxiety and stress related to the pandemic were significantly associated with higher phishing susceptibility, providing empirical evidence that emotionally charged contexts influence user behaviour and decision making during social engineering attacks (Abroshan et al., 2021). Other work on multimodal emotion recognition has highlighted the need to account for cultural variation, subjectivity and ethical constraints when interpreting signals from sound, image and text (Hosseini, Yamaghani and Poorzaker Arabani, 2024). Building on these principles, this thesis applies emotion-aware NLP to smishing by combining message content, emotion cues and user behaviour in a domain-specific way on English SMS data.

User behaviour research further explains how individuals respond to social engineering. Reviews of psychological and cognitive factors indicate that susceptibility to phishing and related cyberattacks is shaped by traits such as cognitive bias, impulsivity and prior experience (Moustafa, Bello and Maurushat, 2021). Real world smishing experiments with 265 participants have found that 16.9 % of recipients engaged with initial malicious messages by clicking links, replying or calling, and that 12.8 % responded again during a follow up wave, providing concrete behavioural evidence of patterned vulnerability rather than random user error (Rahman et al., 2022).

Field studies in workplace environments add further nuance. One large study found that spear phishing emails using authority and urgency cues produced significantly higher click rates than emails without those cues, with authority having the stronger effect (Williams, Hinds and Joinson, 2018). Interpreted through Prospect Theory, these results illustrate how attackers exploit fear of loss and compliance pressure through loss aversion, the certainty effect and framing (Kahneman and Tversky, 1979). For this project, Prospect Theory underpins the design of user personas that vary in sensitivity to reward, urgency and caution.

Additional behavioural evidence comes from simulated phishing campaigns where employees who clicked reported higher immediate stress and lower phishing self-efficacy than those who reported the email (Schöps et al., 2024). Participants generally viewed the campaign as effective, but the study highlights measurable human factors such as stress and self-efficacy that shape susceptibility. These findings support the persona-based simulation layer in this research by justifying how model outputs can be translated into risk perceptions aligned with user archetypes, grounded in observed differences in stress and self-belief rather than untested assumptions.

2.3 Research Gap and Justification

Progress in smishing detection has been driven by advances in NLP and deep learning, yet major gaps remain in emotional sensitivity, behavioural awareness and data diversity. Transformer models such as BERT and DistilBERT have achieved exceptional performance on a range of NLP tasks (Devlin *et al.*, 2019), and recent smishing studies report strong results using contextual embeddings and attention mechanisms (Hosseinpour and Das, 2025; Shen et al., 2025). However, these works also show that performance is highly sensitive to dataset composition, including class balance, curation and domain. Plain transformer classifiers tend to prioritise lexical and semantic patterns rather than the psychological persuasion tactics that characterise social engineering. Emotion aware NLP has progressed rapidly in general domains (Chutia and Baruah, 2024; Xue et al., 2025), but applications to adversarial or deceptive messaging remain limited. At the same time, user behaviour research shows consistent susceptibility patterns in smishing (Rahman et al., 2022; Timko and Rahman, 2024), yet most detection work remains model centric with little incorporation of user feedback or behavioural modelling.

Dataset constraints amplify these issues. The SMS Spam Collection and newer corpora such as SmishTank and SpamDam remain relatively small, English only and irregularly updated (Almeida, Hidalgo and Yamakami, 2011; Li et al., 2024; Timko and Rahman, 2024). Recent work has also shown that generative AI chatbots can be jailbroken to produce convincing smishing content, raising concerns about future attack diversity (Shibli, Pritom and Gupta, 2024). At the same time, persistent challenges with code switched and multilingual data highlight the need for adaptive approaches that go beyond monolingual benchmarks (Doğruöz, Sitaram and Yong, 2023). Taken together, these findings suggest that high benchmark scores on narrow datasets may overstate real world robustness.

Therefore, the contribution of this thesis is to bridge these silos by developing a unified transformer based smishing detection system that integrates emotion analysis and simulated user behaviour. By merging multiple SMS datasets, applying consistent normalisation and combining linguistic, emotional and behavioural indicators, the project aims to produce a model that is not only accurate but also interpretable. The approach is designed to support real world applicability in telecom and mobile security environments by explaining why messages are risky and how different users might react, thereby advancing both technical detection and human centric understanding of smishing threats. The next chapter outlines the research methodology and system design used to implement and evaluate this integrated framework.

3. Research Methodology

This chapter outlines the design, training, and evaluation process for the smishing detection system. The artefact combines natural language processing, emotion analysis, and simulated user behaviour to create an interpretable and practical detection pipeline. Figure 3.1 shows a high-level diagram of the proposed smishing detection framework.

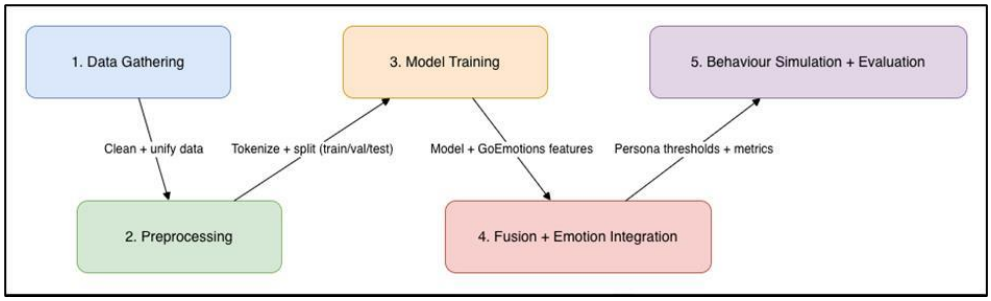

Figure 3.1 – High-level diagram of the proposed smishing detection framework

3.1 Research Design

The project follows a supervised learning approach using labelled SMS messages to train and compare multiple models. The methodology proceeds in three incremental phases. First, lexical baselines are established using TF-IDF features with Logistic Regression, Linear SVC, and Multinomial Naive Bayes. Second, contextual transformer models such as BERT and DistilBERT are fine tuned to capture intent, tone, and meaning. Finally, emotion and behaviour fusion layers are added to integrate semantic outputs with emotion probabilities and personabased interpretation.

A consistent preprocessing pipeline is applied to all datasets, performing Unicode NFKC normalisation, lowercasing, whitespace compression, and replacement of URLs, emails, and numbers with consistent placeholder tokens. Non-English and duplicate messages are removed to ensure data uniformity, and keeping only English messages makes it easier to manually check that predictions and emotions look sensible. This guarantees that TF-IDF vectorizers and transformer tokenizers operate on identical clean text.

The dataset is balanced during training to ensure equal representation of legitimate and smishing samples, while the validation and test sets remain imbalanced to simulate real world message distributions. Evaluation focuses on F1 and PR AUC metrics, since false negatives correspond to missed attacks and false positives contribute to alert fatigue. The overall design is data driven, demonstrating whether emotional and behavioural layers can deliver measurable and explainable improvements beyond lexical and semantic baselines.

3.2 Materials and Environment

All experiments were run in Python using a standard ML and NLP stack, including PyTorch, Hugging Face Transformers, scikit-learn, pandas and numpy amongst others. The full codebase and data pipeline can be cloned from GitHub at https://github.com/Joshoneill6649/smishingnlp-josh-oneill. The configuration manual, README and requirements.txt together provide the exact hardware, library details and script running order needed to reproduce all results end to end.

3.3 Dataset Preparation

Four public SMS datasets were combined to create a unified corpus of legitimate and malicious messages. These were the UCI SMS Spam Collection, SmishTank, SpamDam and the NUS SMS Corpus. Together they mix older and newer campaigns and are widely used in prior SMS spam research, which makes the results easier to compare. Each record was mapped to a shared schema containing text, label, and source fields. Cleaning steps included lowercasing, token

replacement, English filtering, and deduplication. The processed and merged dataset was stored as a combined Parquet file used for all subsequent training and evaluation, with removed rows preserved separately for transparency. Figure 3.2 summarises the dataset composition by source and class before and after cleaning.

source	HAM	SMISH
nus_sms	55801	0
smishtank	0	1055
spamdam	0	76577
uci_sms	4825	747
TOTAL	60626	78379

source	HAM	SMISH
nus_sms	44488	0
smishtank	0	999
spamdam	0	36404
uci_sms	4515	627
TOTAL	49003	38030

Figure 3.2 – Dataset composition by source and class before (left) and after cleaning (right) 3.4

Data Splitting

The combined dataset was divided using a stratified splitting procedure that preserved the ham and smish proportions in each split. Training, validation, and test sets follow an 80/10/10 ratio so that most data are used for learning while still keeping separate partitions for tuning and unbiased evaluation. A balanced training set containing equal numbers of legitimate and smishing messages was also generated, along with validation and test deploy splits that maintain a realistic imbalance of roughly 85% legitimate and 15% smishing messages. This structure provides a strong learning base while also reflecting that real SMS traffic is dominated by legitimate messages. The sizes and class ratios of all splits are summarised in Figure 3.3.

split	N	ham	smish	smish_percent
train	69626	39202	30424	43.7
val	8703	4900	3803	43.7
test	8704	4901	3803	43.69
train_balanced	60848	30424	30424	50
val_deploy	5770	4900	870	15.08
test_deploy	5771	4901	870	15.08

Figure 3.3 – Overview of dataset splitting process.

3.5 Model Development

Each model was trained, validated and tested using identical data splits, preprocessing steps and evaluation metrics. Training used the balanced dataset, while threshold tuning was performed on the deployment style validation split by maximising the F1 score. Final testing used both balanced and imbalanced splits to compare robustness under operational conditions. Baseline TF-IDF models used Logistic Regression, Linear SVC and Multinomial Naive Bayes, providing interpretable lexical benchmarks based on word frequency rather than context. Contextual transformer models (BERT base uncased and DistilBERT base uncased, loaded via the Hugging Face Transformers library) were then fine-tuned with short training schedules and class weighted cross entropy loss to capture nuances such as urgency, authority and persuasion. These transformer models served both as standalone smishing classifiers and as backbones for the later fusion experiments.

3.6 Emotion Aware and Fusion Models

To incorporate affective information, emotional features were extracted using the GoEmotions student model, a DistilBERT variant trained on a large Reddit corpus covering 27 emotion categories plus neutral. The model was used as a fixed feature extractor that produced a probability distribution for each emotion. Two configurations were used in this project: emotion fused with BERT and emotion fused with TF-IDF. In both fusion models the emotion

probability vector was concatenated with the backbone smish probability and passed into a Logistic Regression classifier. The backbone models remained frozen to isolate the contribution of emotion signals, and thresholds were optimised on the validation deployment split using the same F1 based method as before. Figure 3.4 illustrates the late-fusion pipeline used to combine emotion probabilities with backbone smish scores.

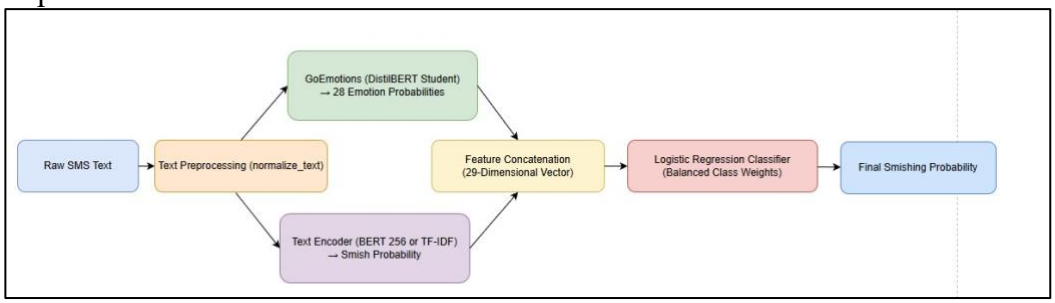

Figure 3.4 - Emotion aware late fusion architecture for the smishing detection system

3.7 Persona Based Behaviour Simulation

A lightweight rule-based persona layer was added to interpret model predictions in human terms. Three simulated user types were defined: a Reward Seeking User, a Busy Professional, and a Cautious User. Given a smish probability and dominant emotion profile, each persona produced an adjusted risk score, a categorical level (Low, Medium, or High), and a brief textual explanation. The logic is transparent, deterministic, and does not involve personal data. Its purpose is to make automated outputs more relatable and interpretable to end users. The persona definitions and emotion sensitivities are summarised in Figure 3.5.

Persona	Description	Sensitive to Increase Risk	Protective (Decrease Risk)	Base Click Risk
Busy Professional	High message volume and time pressure; more likely to react quickly to urgent or threatening messages without verifying.	fear, anxiety, urgency, embarrassment, anger	calm, neutral	0.35
Reward-Seeking User	Likes rewards, refunds, prizes, discounts; more responsive to positive and gain-framed emotional cues.	joy, excitement, desire, admiration	fear, disgust	0.30
Cautious User	Security-aware and skeptical; less influenced by emotional tone and more likely to double-check suspicious content.	—	fear, neutral, annoyance	0.10

Figure 3.5 – Persona definitions and emotion sensitivities used in the behaviour simulation

3.8 Ethics and Reproducibility

All datasets used in this study are publicly available SMS corpora that have already been anonymised, so no phone numbers, names or other direct identifiers are processed. The project does not collect any new messages from users, and the emotion and persona components are used only for defensive analysis of incoming texts rather than to design more persuasive attacks. Persona profiles are illustrative and do not include demographic attributes. Reproducibility is supported through the public project repository, which contains the preprocessing and training scripts, a configuration manual with environment and command details, and a structured directory layout with fixed random seeds so that the same data splits, models and evaluation reports can be regenerated from the raw SMS files.

4. Design and Implementation

This chapter describes how the methodology from Chapter three was implemented as a modular and reproducible system. The artefact is structured as a pipeline that moves from raw SMS data through model training and fusion to persona-based risk interpretation and evaluation.

4.1 System Architecture and Workflow

The implementation is organised as a pipeline with distinct stages for data processing, model training, emotion modelling, persona logic and command line interaction. Data preprocessing scripts load the raw SMS corpora, normalise and clean the text, filter non-English messages and write a combined Parquet file to the processed data directory. A splitting script then creates the canonical train, validation, test, train balanced, val_deploy and test_deploy partitions described in Chapter three.

On top of this data stage, the model components train classical TF-IDF baselines and transformer models on the shared splits, evaluate them and select the best TF-IDF and transformer backbones using F1 and PR AUC on the deployment style validation set. Fusion scripts then concatenate the smish probability from the selected backbone model with the GoEmotions probability vector and train a Logistic Regression layer to produce emotion aware models, using two backbones in practice: the strongest transformer (BERT base with a 256 token context) and the strongest TF-IDF baseline (Linear SVC). At inference time the prediction script provides a simple command line interface: it loads the selected model, obtains smish probabilities and emotion vectors for a single message and passes them to the persona module, which generates user specific risk levels and explanations. All components read from the processed data directory, save models in a dedicated models location and write metrics and prediction files to the report's directory for full traceability. Figure 4.1 summarises this end-toend pipeline from raw SMS input to persona level risk scores.

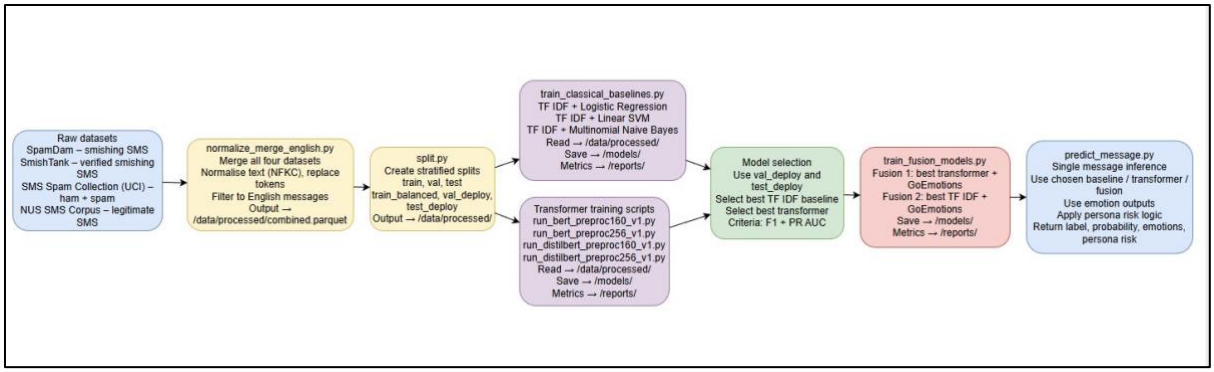

Figure 4.1 – End-to-end architecture of the proposed smishing detection pipeline

4.2 Model Training Configuration

TF-IDF baselines share a single feature and training setup so that differences in performance come from the classifier rather than the vectorizer. A random seed of 1337 is used for all runs. The TF-IDF vectorizer builds unigram and bigram features with a minimum document frequency of 2 and a maximum document frequency of 0.95, which removes extremely rare and very common terms. Logistic Regression and Linear SVC use a regularisation parameter C of 1.0, and the Multinomial Naive Bayes model uses a smoothing parameter alpha of 1.0.

These conservative settings gave stable performance in pilot runs and avoid over tuning individual baselines.

Transformer models are fine-tuned under a single protocol so that the only major differences between runs are the base model and the context length. BERT base uncased and DistilBERT base uncased are trained with seed 1337 for three epochs using AdamW with a learning rate of 2e−5, batch size 16, weight decay 0.01 and a warmup ratio of 0.10. The maximum sequence length is set to either 160 or 256 tokens depending on the experiment, and gradients are clipped at a norm of 1.0 to prevent instability. Training progress is logged every 50 steps and class weighted loss is used to offset remaining class imbalance. Full command line options and environment details are documented in the configuration manual and project repository. These hyperparameters follow standard BERT fine-tuning practice and were chosen to balance performance with training stability on the available hardware.

4.3 Behaviour Logic

Persona behaviour is implemented in a standalone module defining rules for three archetypes. Reward seeking users are influenced by excitement and the promise of gain, busy professionals are sensitive to urgency and authority, and cautious users are conservative but not immune to deception. Each rule set interprets the emotional tone and model output to generate a risk level and explanation. The simplicity and transparency of this module ensure that analysts can audit or modify the reasoning process easily.

4.4 Outputs and Artefacts

The completed artefact generates structured outputs, including processed datasets, model checkpoints, prediction files, and evaluation reports. This separation of components facilitates reproducibility and allows researchers to retrain or extend the system on new data without altering its architecture.

In summary, the implementation realises the research methodology as a modular and transparent artefact. It integrates lexical, contextual and emotional reasoning, producing interpretable risk assessments that can be reproduced and verified. The next chapter evaluates these models quantitatively and qualitatively to determine how well the implementation meets the research objectives.

5. Evaluation

This chapter evaluates the proposed smishing detection framework under both balanced and deployment-style conditions. It outlines the experimental setup, compares baseline and transformer models, analyses the impact of emotion fusion, and uses persona-based examples to illustrate how predictions translate into human-interpretable risk assessments.

5.1 Experimental Setup

Two evaluation regimes were applied to measure performance under both balanced and realistic conditions. The first uses a 50:50 distribution of legitimate and smishing messages to assess discriminative ability, while the second reflects real world imbalance with approximately 15%

smishing content. Thresholds were fixed after optimisation on the validation set to ensure fair comparison and prevent data leakage.

5.2 Overall Model Comparison

Transformer based models consistently outperformed lexical baselines across all evaluation metrics. The BERT model with a 256 token context achieved an F1 score of 0.9874 on the deployment style test, maintaining both high precision .9829 and recall .9920. Classical TFIDF models achieved strong results ranging from F1 0.9644 to 0.9786 but were slightly less robust under imbalance. The narrow gap between balanced and imbalanced testing indicates good generalisation and minimal overfitting. These results for all baseline, transformer, and emotion fusion models are summarised in Figure 5.1.

Model	F1 (test)	F1 (test_deploy)	Precision (deploy)	Recall (deploy)	ROC-AUC (deploy)	PR-AUC (deploy)	Thr (deploy)
bert_preproc256_v1	0.9933	0.9874	0.9829	0.9920	0.9987	0.9976	0.9933
emotion_fusion_model_1	0.9938	0.9846	0.9773	0.9920	0.9977	0.9897	0.9989
bert	0.9936	0.9835	0.9751	0.9920	0.9989	0.9979	0.9898
distilbert_preproc256_v1	0.9929	0.9812	0.9708	0.9920	0.9995	0.9978	0.9704
distilbert	0.9933	0.9807	0.9707	0.9908	0.9994	0.9978	0.5000
tfidf_linearsvc	0.9786	0.9706	0.9745	0.9667	0.9963	0.9805	0.5479
emotion_fusion_tfidf_lsvc_1	0.9749	0.9682	0.9733	0.9632	0.9960	0.9832	0.9963
tfidf_lr	0.9743	0.9609	0.9620	0.9598	0.9956	0.9777	0.5935
tfidf_mnb	0.9644	0.9473	0.9654	0.9299	0.9925	0.9792	0.6285

Figure 5.1 – Comparison scores for TF-IDF, Transformer, and Emotion fusion models.

5.3 Baselines versus Transformers

The comparison confirms that while TF-IDF provides an effective baseline, transformer models achieve superior contextual understanding. The shorter transformer runs used a 160 token window, which sometimes truncated URLs and the tail end of longer scam narratives. In contrast, the 256 token BERT model preserved full URLs and longer scam phrases, allowing it to capture more complete linguistic cues linked to social engineering tactics.

5.4 Effect of Emotion Fusion

Late fusion results showed that incorporating emotion features produces performance that is broadly competitive with the best transformer baseline rather than clearly superior to it. On the balanced test split the emotion fusion model reaches an F1 of 0.9938, just ahead of the 256 token BERT model at 0.9933. However, on the deployment test split the 256 token BERT model still holds the strongest overall scores, with an F1 of 0.9874 and PR AUC of 0.9976 compared with 0.9846 and 0.9897 for the emotion fusion model. The main advantage of emotion information is therefore qualitative rather than purely quantitative. It enables interpretability by surfacing key affective drivers such as fear, urgency, and anticipation, which are strongly associated with manipulation in smishing messages. This transparency helps users understand why messages are flagged and makes the model's decisions easier to trust.

5.5 Qualitative Persona View

Qualitative analysis through the persona layer demonstrated how a single classifier output can be contextualised for different user types. Benign messages expressing gratitude or calm tone produced very low smish probabilities and were rated as low risk by all personas. Conversely, fraudulent messages referencing rewards or refunds triggered high probabilities and were judged as high risk for the Reward Seeking and Busy Professional profiles, and medium risk

for the Cautious profile. This shows that the emotion and persona layers together provide a bridge between technical classification and human behaviour.

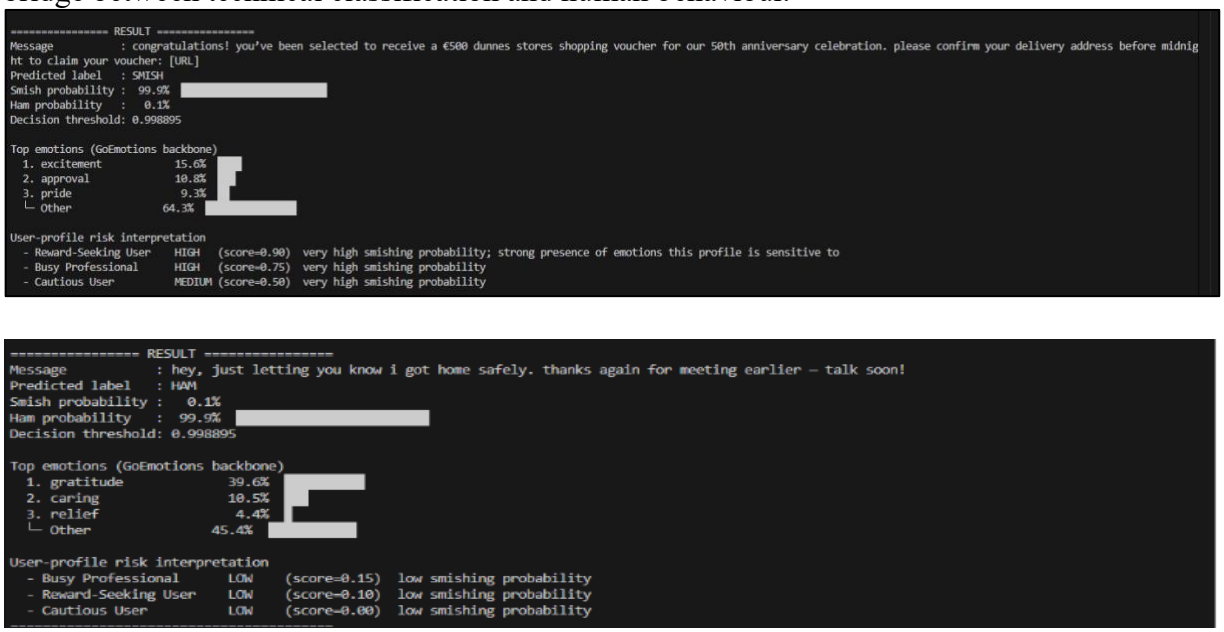

Figure 5.2 - Console outputs for smishing and ham messages showing prediction, emotion profile, and persona-based risk

6. Discussion and Conclusion

This chapter discusses how the evaluation results answer the research question and position the proposed artefact within existing work. It interprets the quantitative findings for baselines, transformers, and emotion fusion, reflects on the role of the persona layer, and then summarises the main contributions, limitations, and directions for future research.

The experiments show that smishing can be detected with very high accuracy while still giving explanations that match how different users experience risk. As shown across all models in Figure 5.1, there is a clear hierarchy. The best BERT configuration with a 256 token context reaches an F1 score of 0.9874 on the deployment style split, with precision 0.9829 and recall 0.9920, so fewer than ten malicious messages in a thousand would be missed. Classical TFIDF baselines remain strong, with deployment F1 scores between 0.9473 and 0.9706, but they still underperform the transformer models across all reported metrics (Figure 5.1).

This pattern confirms the first part of the research argument. TF-IDF with Linear SVC or Logistic Regression provides a solid reference point and reflects earlier SMS spam work, but transformer models deliver extra performance on the combined corpus. The best BERT model improves F1 by roughly one to two percentage points over the strongest TF-IDF baseline on the deployment split and achieves a higher precision recall area. These gains are modest but important in a security setting where each extra percentage point of recall represents fewer successful smishing attempts. The small difference between scores on the balanced and deployment regimes for the top models suggests that the split design and threshold tuning strategy generalise rather than simply overfit training data. Within the transformer family, context length is a key factor: runs with 160 tokens sometimes truncated URLs and the tail of

longer narratives, while the 256 token BERT model preserved these segments and achieved the strongest scores.

The emotion fusion experiments address the second part of the research question. Numerically, the late fusion model that combines BERT with GoEmotions probabilities is competitive with the pure BERT baseline rather than clearly superior. On the balanced test split the fusion model achieves an F1 of 0.9938 compared with 0.9933 for BERT, while on the deployment split it falls slightly behind with an F1 of 0.9846 and precision recall area of 0.9897 versus 0.9874 and 0.9976. This suggests that the contextual language model already captures most of the discriminative signal and that emotion features add limited extra power under realistic imbalance. The main value of emotion is therefore qualitative. By converting each SMS into a distribution over emotions such as excitement, fear or gratitude, the system surfaces the affective tactics that drive a decision.

The voucher scam and follow up examples in Figure 5.2 illustrate this effect. In the scam message the model assigns a very high smishing probability and highlights excitement and approval as dominant emotions, which the persona layer translates into a high-risk rating for Reward Seeking and Busy Professional profiles and a medium rating for the Cautious profile. In the benign follow up the system emphasises gratitude and caring, assigns a very low smishing probability and rates all personas as low risk. These examples show that the emotion and persona layers together provide a bridge between technical classification and user perception, which was a core motivation for the project.

Overall, the results answer the research question as follows. Transformer models provide a measurable improvement over TF-IDF baselines for smishing detection on a combined corpus that merges four public datasets. Emotion features, when added as a late fusion layer, maintain this high level of performance while enabling richer explanations about why a message is considered risky. The persona-based behaviour simulation shows that a single classifier output can be reframed for different user types without retraining, which is relevant for telecom providers or security teams who support users with different risk tolerances and levels of security awareness.

These findings sit within several limitations. All experiments are conducted on English only subsets of public SMS corpora, because non-English messages were filtered out during preprocessing. As a result, the system may not fully capture regional, multilingual, or newly emerging smishing campaigns. The emotion model is trained on Reddit comments rather than SMS, which may misinterpret very short or heavily abbreviated messages. Fusion relies on frozen backbones and a simple Logistic Regression layer, exploring only one part of the design space for combining language and emotion signals. Persona definitions are informed by prior work on phishing and user behaviour but are not validated through new user studies. For these reasons the artefact should be viewed as a research prototype that shows the potential of combining language, emotion, and behaviour rather than as a production ready tool.

Despite these constraints the project delivers several contributions. It constructs a unified smishing benchmark by merging multiple datasets under a consistent preprocessing pipeline, shows that carefully tuned transformer models can achieve near perfect detection rates even under class imbalance, and introduces an emotion and persona layer that turns classifier outputs into more human centred risk statements. It also provides a documented codebase and configuration manual so that the full pipeline can be reproduced or extended on new data.

Future work can build on this foundation by collecting more recent and diverse smishing messages, exploring joint training of language and emotion components, and testing robustness against obfuscation, adversarial examples and multilingual content. Behavioural validation of the persona logic and integration with real security workflows, such as analyst dashboards or on device warnings, would further enhance the practical value of the approach. Taken together, the results indicate that combining contextual language modelling with emotion analysis and simple behavioural rules is a promising direction for making smishing detection both accurate and aligned with how people read and react to messages.

References

Devlin, J. *et al.* (2019) "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in J. Burstein, C. Doran, and T. Solorio (eds.) *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4171–4186. Available at: https://doi.org/10.18653/v1/N19-1423.

Tanbhir, G. *et al.* (2024) "Hybrid Machine Learning Model for Detecting Bangla Smishing Text Using BERT and Character-Level CNN," in *2024 13th International Conference on Electrical and Computer Engineering (ICECE)*, pp. 57–62. Available at: https://doi.org/10.1109/ICECE64886.2024.11024872.

Abroshan, H. *et al.* (2021) "COVID-19 and Phishing: Effects of Human Emotions, Behavior, and Demographics on the Success of Phishing Attempts During the Pandemic," *IEEE Access*, 9, pp. 121916–121929. Available at: https://doi.org/10.1109/ACCESS.2021.3109091.

Almeida, T.A., Hidalgo, J.M.G. and Yamakami, A. (2011) "Contributions to the study of SMS spam filtering: new collection and results," in *Proceedings of the 11th ACM symposium on Document engineering*. New York, NY, USA: Association for Computing Machinery (DocEng '11), pp. 259–262. Available at: https://doi.org/10.1145/2034691.2034742.

Chutia, T. and Baruah, N. (2024) "A review on emotion detection by using deep learning techniques," *Artificial Intelligence Review*, 57(8), p. 203. Available at: https://doi.org/10.1007/s10462-024-10831-1.

Dobson, R. *et al.* (2024) "Don't Forget the Humble Text Message: 25 Years of Text Messaging in Health," *Journal of Medical Internet Research*, 26, p. e59888. Available at: https://doi.org/10.2196/59888.

Doğruöz, A.S., Sitaram, S. and Yong, Z.X. (2023) "Representativeness as a Forgotten Lesson for Multilingual and Code-switched Data Collection and Preparation," in H. Bouamor, J. Pino, and K. Bali (eds.) *Findings of the Association for Computational Linguistics: EMNLP 2023*. *Findings 2023*, Singapore: Association for Computational Linguistics, pp. 5751–5767. Available at: https://doi.org/10.18653/v1/2023.findings-emnlp.382.

El-Haj, M. and Takanami, R. (2023) "Unifying Emotion Analysis Datasets using Valence Arousal Dominance (VAD)," in S. Carvalho et al. (eds.) *Proceedings of the 4th Conference on Language, Data and Knowledge*. *LDK 2023*, Vienna, Austria: NOVA CLUNL, Portugal, pp. 220–225. Available at: https://aclanthology.org/2023.ldk-1.19/ (Accessed: November 11, 2025).

ENISA Threat Landscape 2024 | ENISA (2024). Available at: https://www.enisa.europa.eu/publications/enisa-threat-landscape-2024 (Accessed: September 18, 2025).

Hosseini, S., Yamaghani, M.R. and Poorzaker Arabani, S. (2024) "A review of the methods of recognition multimodal emotions in sound, image and text." Available at: https://doi.org/10.71885/IJORLU-2024-1-657.

Hosseinpour, S. and Das, S. (2025) "POSTER: A Multi-Signal Model for Detecting Evasive Smishing." arXiv. Available at: https://doi.org/10.48550/arXiv.2505.18233.

Jamal, S., Wimmer, H. and Sarker, I.H. (2024) "An improved transformer-based model for detecting phishing, spam and ham emails: A large language model approach," *SECURITY AND PRIVACY*, 7(5), p. e402. Available at: https://doi.org/10.1002/spy2.402.

Kahneman, D. and Tversky, A. (1979) "Prospect Theory: An Analysis of Decision under Risk," *Econometrica*, 47(2), pp. 263–291. Available at: https://doi.org/10.2307/1914185.

Laureano De Leon, F.A. *et al.* (2025) "UoB-NLP at SemEval-2025 Task 11: Leveraging Adapters for Multilingual and Cross-Lingual Emotion Detection," in S. Rosenthal et al. (eds.) *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*. *SemEval 2025*, Vienna, Austria: Association for Computational Linguistics, pp. 1570–1576. Available at: https://aclanthology.org/2025.semeval-1.207/ (Accessed: August 9, 2025).

Li, Y. *et al.* (2024) "SpamDam: Towards Privacy-Preserving and Adversary-Resistant SMS Spam Detection." arXiv. Available at: https://doi.org/10.48550/arXiv.2404.09481.

Machová, K. *et al.* (2023) "Detection of emotion by text analysis using machine learning," *Frontiers in Psychology*, 14. Available at: https://doi.org/10.3389/fpsyg.2023.1190326.

Mahmud, T. *et al.* (2024) “Enhancing Cybersecurity: Hybrid Deep Learning Approaches to Smishing Attack Detection,” *Systems*, 12(11), p. 490. Available at: https://doi.org/10.3390/systems12110490.

Moustafa, A.A., Bello, A. and Maurushat, A. (2021) “The Role of User Behaviour in Improving Cyber Security Management,” *Frontiers in Psychology*, 12. Available at: https://doi.org/10.3389/fpsyg.2021.561011.

Pourmohamad, R. *et al.* (2024) “Deep Dive into Client-Side Anti-Phishing: A Longitudinal Study Bridging Academia and Industry,” in *Proceedings of the 19th ACM Asia Conference on Computer and Communications Security*. New York, NY, USA: Association for Computing Machinery (ASIA CCS ’24), pp. 638–653. Available at: https://doi.org/10.1145/3634737.3657027.

Rahman, M.L. *et al.* (2022) “Users really do respond to smishing.” arXiv. Available at: https://doi.org/10.48550/arXiv.2212.13312.

Saputra, N. (2024) “Analysis of SMS Spam Detection using Tf-Idf: A Study On SMS Spam Collection Dataset,” *Jurnal Sosial Teknologi*, 4, pp. 213–217. Available at: https://doi.org/10.59188/jurnalsostech.v4i4.1214.

Scam Alerts – Stay Vigilant | An Post (2024). Available at: https://www.anpost.com/Media-Centre/News/Scam-Alerts-%E2%80%93-Stay-Vigilant (Accessed: July 28, 2025).

Schöps, M. *et al.* (2024) “Simulated Stress: A Case Study of the Effects of a Simulated Phishing Campaign on Employees’ Perception, Stress and Self-Efficacy,” in. *33rd USENIX Security Symposium (USENIX Security 24)*, pp. 4589–4606. Available at: https://www.usenix.org/conference/usenixsecurity24/presentation/schops (Accessed: September 7, 2025).

Shen, L. *et al.* (2025) “SMS spam detection using BERT and multi-graph convolutional networks,” *International Journal of Intelligent Networks*, 6, pp. 79–88. Available at: https://doi.org/10.1016/j.ijin.2025.06.002.

Shibli, A.M., Pritom, M.M.A. and Gupta, M. (2024) “AbuseGPT: Abuse of Generative AI ChatBots to Create Smishing Campaigns.” arXiv. Available at: https://doi.org/10.48550/arXiv.2402.09728.

Tabassum, S., Faklaris, C. and Lipford, H.R. (2024) “What Drives SMiShing Susceptibility? A U.S. Interview Study of How and Why Mobile Phone Users Judge Text Messages to be Real or Fake,” in. *Twentieth Symposium on Usable Privacy and Security (SOUPS 2024)*, pp. 393–411. Available at: https://www.usenix.org/conference/soups2024/presentation/tabassum-sarah (Accessed: September 20, 2025).

Timko, D., Castillo, D.H. and Rahman, M.L. (2023) *A Quantitative Study of SMS Phishing Detection*, *arXiv.org*. Available at: https://arxiv.org/abs/2311.06911v4 (Accessed: October 2, 2025).

Timko, D. and Rahman, M.L. (2024) “Smishing Dataset I: Phishing SMS Dataset from Smishtank.com,” in *Proceedings of the Fourteenth ACM Conference on Data and Application Security and Privacy*, pp. 289–294. Available at: https://doi.org/10.1145/3626232.3653282.

“Verizon” (2025). Available at: https://www.verizon.com/business/resources/reports/2025-dbir-executive-summary.pdf

Williams, E.J., Hinds, J. and Joinson, A.N. (2018) “Exploring susceptibility to phishing in the workplace,” *International Journal of Human-Computer Studies*, 120, pp. 1–13. Available at: https://doi.org/10.1016/j.ijhcs.2018.06.004.

Xue, J. *et al.* (2025) “JNLP at SemEval-2025 Task 11: Cross-Lingual Multi-Label Emotion Detection Using Generative Models,” in S. Rosenthal et al. (eds.) *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*. Vienna, Austria: Association for Computational Linguistics, pp. 20–27. Available at: https://aclanthology.org/2025.semeval-1.4/.