University of Southampton

Faculty of Engineering, Science and Mathematics
School of Electronics and Computer Science

An NLP-Driven Framework for Business Email Compromise Detection and Authorship Verification

by

Amirah M Almutairi, MSc, SFHEA

Supervisors: Dr. Nawfal Al Hashimy and Dr. BooJoong Kang

ORCiD: 0000-0002-2194-7936

A thesis for the degree of
Doctor of Philosophy

September 2025

University of Southampton

Abstract

Faculty of Engineering, Science and Mathematics
School of Electronics and Computer Science

Doctor of Philosophy

An NLP-Driven Framework for Business Email Compromise Detection and Authorship Verification

by Amirah M Almutairi, MSc, SFHEA

Business Email Compromise (BEC) presents a critical cybersecurity threat, leveraging linguistic impersonation and social engineering rather than traditional malicious payloads. These attacks routinely evade conventional filters by mimicking legitimate communication styles and exploiting trusted identities.

This thesis explores content-based detection strategies for BEC using a sequence of natural language processing (NLP) models. First, it proposes a transformer-based classifier to detect semantic indicators of deception in email body text. Second, it develops a Siamese authorship verification (AV) model that captures stylistic consistency, even under adversarial mimicry. These components are unified within a multi-task learning (MTL) framework that simultaneously optimizes for BEC detection and AV by sharing underlying representations while preserving task-specific objectives.

To support empirical evaluation, a structured taxonomy of BEC fraud is introduced, and a synthetic email dataset is generated through prompt-guided language model fine-tuning and human validation. Experiments on combined real and synthetic corpora demonstrate that the MTL model achieves up to 97% F1-score in BEC detection and 93% in AV, outperforming transfer learning baseline while reducing false positives and computational overhead.

This work contributes a principled, modular, and extensible framework for enhancing email security through joint semantic and stylistic analysis, addressing gaps in current defenses against sophisticated impersonation attacks.

Contents

List of Figures

List of Tables

Declaration of Authorship

I declare that this thesis and the work presented in it is my own and has been generated by me as the result of my own original research.

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
7. Parts of this work have been published as:
   (a) Almutairi, A. M., Kang, B., & Al Hashimy, N. "Business Email Compromise: A Systematic Review of Understanding, Detection, and Challenges." *Computers & Security*, 2025. doi:10.1016/j.cose.2025.104630
   (b) Almutairi, A. M., Kang, B., & Al Hashimy, N. "Business Email Compromise: A Comprehensive Taxonomy for Detection and Prevention." *Proc. 7th Int. Conf. on Information Science and Systems (ICISS '24)*, 2024, pp. 49–54. doi:10.1145/3700706.3700714
   (c) Almutairi, A., Kang, B., & Fadhel, N. "The Effectiveness of Transformer-Based Models for BEC Attack Detection." *Network and System Security (NSS 2023)*, LNCS 13983, 2023. doi:10.1007/978-3-031-39828-5_5
   (d) Almutairi, A., Kang, B., & Al Hashimy, N. "BiBERT-AV: Enhancing Authorship Verification Through Siamese Networks with Pre-trained BERT and Bi-LSTM." *Int. Conf. on Ubiquitous Security (UDSec 2023)*, LNCS 13984, 2023. doi:10.1007/978-981-97-1274-8_2

Acknowledgements

All praise and gratitude to Allah, who grants strength in moments of weakness and clarity in times of doubt. This PhD journey has been one of striving, learning, and growing, an achievement that stands not only on my efforts but on the shoulders of those who walked beside me, and lifted me through it all.

To my beloved father, my dearest mother, and my precious brother Sultan, may Allah have mercy on their souls. Your memory has been my anchor and my motivation. I carried your names in every milestone, your love in every word, and your strength in every hardship. This achievement is, in every sense, a tribute to you.

To my husband Fahad, your love, patience, and quiet strength have been my shelter. Thank you for enduring the long nights, the absences, the stress, and the silence, while always offering me peace and belief. You have been the heart of my resilience.

To my children, Hoor, Sultan, Malak, Shaihan, and Rakan, you are my light and my greatest source of purpose. I hope this journey teaches you the value of persistence, faith, and holding onto your dreams.

To my supervisor, Dr. Nawfal, your guidance went far beyond academic support. Your confidence in my ideas, your thoughtful feedback, and your consistent words; “Well done, I’m proud of you”, carried me more than you may know. Thank you for shaping both the path and the person.

To my supervisor, Dr. BooJoong, thank you for every “why” that pushed me deeper, for the challenges that sharpened my thinking, and for the trust you placed in my work.

I am also deeply grateful to my entire family, my brothers and sisters, for their enduring love and support. To my dear friends, thank you for your unwavering friendship, prayers, and constant check-ins. I extend heartfelt thanks to the Government of Saudi Arabia and Shaqra University for fully funding four years of my doctoral studies.

This thesis is more than research it is the story of love, loss, belief, and perseverance. I dedicate it to those I carry in my heart and to those who stood by me—always.

To Saudi Arabia, my homeland, my pride, and my inspiration.

Abbreviations

AdamW	Adam Optimization with Weight Decay Regularization
AUC	Area Under the ROC Curve
AV	Authorship Verification
BEC	Business Email Compromise
BiBERT	BERT and BiLSTM-based Siamese Network for Authorship Verification
BiLSTM	Bidirectional Long Short-Term Memory
BERT	Bidirectional Encoder Representations from Transformers
BLEU	Bilingual Evaluation Understudy Score
CAPE	Context-Aware Phishing Email Detection
CEAS	Conference on Email and Anti-Spam Dataset
DL	Deep Learning
DistilBERT	A lightweight version of BERT
DKIM	DomainKeys Identified Mail
GloVe	Global Vectors for Word Representation
GNN	Graph Neural Network
IP	Internet Protocol
LLM	Large Language Model
LSTM	Long Short-Term Memory
MiTM	Man-in-The-Middle Attack
ML	Machine Learning
MTL	Multi-Task Learning
NLP	Natural Language Processing
PAN	Plagiarism, Authorship, and Near-Duplicate Detection Evaluation
QA	Question-Answering
ROC	Receiver Operating Characteristic
ROUGE	Recall-Oriented Understudy for Gisting Evaluation
RNN	Recurrent Neural Network
SHAP	SHapley Additive exPlanations
Siamese	A neural network architecture for similarity learning
SPF	Sender Policy Framework
$TF-IDF$	Term Frequency-Inverse Document Frequency
TREC	Text REtrieval Conference Dataset

Chapter 1

Introduction

1.1 Motivation

Business Email Compromise (BEC) is a targeted form of cyber fraud in which adversaries exploit human trust—often through linguistic impersonation and social engineering—to deceive organisations into transferring funds or divulging sensitive information Federal Bureau of Investigation (2024). Unlike conventional phishing, BEC emails typically avoid overt indicators such as malicious links or attachments, making them particularly difficult to detect using traditional spam filters. High-profile incidents, including those affecting companies like Google and Facebook, have demonstrated the financial and reputational consequences of such attacks, with losses exceeding tens of millions of dollars Internet Crime Complaint Center (IC3) (2023). Given that global email volumes surpass 392 billion messages daily across more than 4.8 billion users Statista (2024), even a modest failure rate poses significant operational and financial risk to enterprises.

Most commercial email security solutions rely on metadata-level features to identify potentially malicious messages. These include sender IP addresses, which identify the originating mail server; authentication protocols like Sender Policy Framework (SPF) and DomainKeys Identified Mail (DKIM), which verify that messages come from authorised sources; and domain reputation scores based on prior behaviour. Such features are generally effective for detecting spam or phishing attempts originating from unauthorised domains. However, in adversarial scenarios where a legitimate account has been compromised, these metadata-based indicators often remain unaltered. The attacker may send emails from trusted infrastructure, bearing valid headers and domain credentials. In such cases, metadata-based checks offer little or no indication of compromise, particularly when the message body contains no overt phishing cues. Therefore, metadata alone is insufficient for detecting advanced threats such as BEC, especially when attackers operate within the bounds of legitimate email systems.

In contrast, the email body is essential to the communication itself and cannot be obscured without losing meaning. It conveys both the semantic content of the request and the stylistic

patterns of the sender. This motivates the exploration of content-based detection methods that focus on linguistic features rather than metadata-level features. Natural Language Processing (NLP) provides a foundation for such approaches, offering methods for modelling semantic and syntactic features in email body content.

Yet semantic inspection alone may not suffice. When attackers convincingly mimic the tone, vocabulary, and formatting of trusted individuals, the email can appear contextually appropriate and bypass traditional filters. To address this limitation, it is necessary to consider not only what is said but how it is expressed. Stylometric research suggests that individual writing styles exhibit stable lexical and syntactic characteristics over time across varied topics and contexts Mendenhall (1887); Bagavandas and Manimannan (2008); Wang (2007). Detecting deviations from these habits can reveal subtle forms of impersonation. As such, AV offers a complementary content-based signal that may enhance the detection of sophisticated BEC attacks Stamatatos (2009); Koppel et al. (2011).

1.2 Research Problem

Although recent NLP-based approaches have improved the semantic analysis of emails Gascon et al. (2018); Cidon et al. (2019), they primarily rely on observable phishing cues—such as malicious links or urgency-related keywords—and often overlook whether the message aligns with the known writing style of the sender. This limitation is particularly problematic in (BEC), where attackers often send well-crafted, natural-language messages from already compromised accounts. In such cases, conventional defenses based on metadata or superficial content patterns may fail, as the messages appear legitimate in both structure and context.

Detecting BEC under these conditions requires examining not only *what* is being said (semantic content) but also *how* it is said (stylistic expression). Semantic-based detection focuses on intent indicators like requests for funds or changes in behaviour. In contrast, stylistic-based detection examines writing style—lexical choices, sentence structure, and syntactic patterns—based on the assumption that legitimate users write in consistent ways. When attackers imitate this style, subtle deviations may still be detectable. This thesis addresses the gap by proposing a unified framework that jointly models semantic anomalies and stylistic inconsistency for improved BEC detection in the early stage.

1.3 Research Aim and Objectives

Aim. This research aims to detect (BEC) and verify authorship using the semantic and stylistic content of email body content. It investigates two core tasks: identifying deceptive intent through semantic analysis, and validating sender identity through writing-style consistency. To address both tasks jointly, this thesis proposes a unified content-based framework leveraging multi-task

learning (MTL), enabling effective detection even in scenarios where metadata-derived features are absent or uninformative.

Objectives.

1. Conduct a systematic literature review of BEC, including common attack strategies and defence mechanisms.
2. Propose a semantic classifier based on transformer models and benchmark it against lexical and heuristic baselines.
3. Propose a stylistic verification module to assess author consistency using deep learning.
4. Integrate both components into an MTL architecture and evaluate performance under varied attack scenarios.

1.4 Scope

This thesis investigates the viability of detecting (BEC) attacks by analyzing the semantic and stylistic properties of English-language email body content. The approach intentionally omits metadata-based features—such as authentication headers (e.g., SPF, DKIM), sender IP addresses, and routing traces—to isolate the predictive capacity of linguistic signals and evaluate their contribution as a standalone detection layer. This design choice reflects practical and adversarial considerations: in compromised-account scenarios, metadata features often remain valid and can therefore mask malicious intent, whereas the body text may carry subtle semantic or stylistic inconsistencies that reveal impersonation attempts.

Moreover, since email content is preserved across archival systems and delivery platforms, it provides a universally accessible, platform-independent input for modeling. By focusing on this layer, the thesis introduces a detection strategy that can operate alongside existing security mechanisms and remain effective in environments where metadata is incomplete, misleading, or adversary-controlled.

1.5 Research Questions

This thesis is guided by one overarching research question:

> **Main Research Question:** *How can NLP-based models be designed to detect Business Email Compromise (BEC) attacks and verify email authorship using only the email body content, while addressing challenges such as impersonation, stylistic mimicry, and lack of metadata?*

To address this central question, the following sub-research questions (SRQs) were formulated:

- **SRQ1:** What are the existing technical and non-technical countermeasures for BEC detection, and what gaps remain?
- **SRQ2:** How can BEC attack strategies be systematically categorized to inform detection design?
- **SRQ3:** How effective are transformer-based classifiers for phishing text-based attacks, and to what extent do they generalise to impersonation-driven BEC when only email body content is available?
- **SRQ4:** Can authorship verification methods based on semantic and stylistic cues reliably distinguish between genuine and impersonated business emails?
- **SRQ5:** Can a unified NLP-based model jointly perform BEC detection and authorship verification through multi-task learning, and how does it compare to sequential or single-task baselines?

1.6 Contributions

1. Presents a comprehensive survey of BEC, including a multi-axis taxonomy that links tactics, adversary goals, and countermeasures.
2. Introduces a transformer-based content detector that outperforms lexical baselines on public benchmarks.
3. Proposes *BiBERT-AV*, a Siamese architecture combining BERT and BiLSTM for stylistic authorship verification.
4. Combine BEC detection and BiBERT-AV models in an MTL setup that jointly detects semantic fraud and stylistic inconsistencies.
5. Provides a thorough empirical evaluation, including false-positive analysis and inference latency.

1.7 Publications

1. Almutairi, A., Kang, B., and Al Hashimy, N. (2025). *Systematic Review on : Understanding, Detection, and Challenges*. doi:10.1016/j.cose.2025.104630. **Chapter3**
2. Almutairi, A. M., Kang, B., & Al Hashimy, N. (2024). *Business Email Compromise: A Comprehensive Taxonomy for Detection and Prevention*. In *Proceedings of the 7th International Conference on Information Science and Systems (ICISS '24)*, pp. 49–54. https://doi.org/10.1145/3700706.3700714 **Chapter 5**

3. Almutairi, A., Kang, B., Fadhel, N. (2023). *The Effectiveness of Transformer-Based Models for BEC Attack Detection*. In: Li, S., Manulis, M., Miyaji, A. (eds) *Network and System Security*. NSS 2023. Lecture Notes in Computer Science, vol 13983. Springer, Cham. https://doi.org/10.1007/978-3-031-39828-5_5 **Chapter 7**

4. Almutairi, A., Kang, B., Al Hashimy, N. (2023). *BiBERT-AV: Enhancing Authorship Verification Through Siamese Networks with Pre-trained BERT and Bi-LSTM*. In: Manulis, M., Miyaji, A., Zhang, Y. (eds) *International Conference on Ubiquitous Security*. Lecture Notes in Computer Science, vol 13984. Springer, Cham. https://doi.org/10.1007/978-3-031-xxxxx-x **Chapter 8**

5. Almutairi, A., Kang, B.,and Al Hashimy, N. (2024). *Integrating Business Email Compromise Detection and Authorship Verification Through Multi-Task Learning*. Submitted and currently *Under Review* at the *Journal of Information Security and Applications*. **Chapter 9**

1.8 Thesis Structure

The remainder of this thesis is structured as follows:

- **Chapter 2: Background** Establishes the foundational concepts required to understand (BEC) and the rationale behind using advanced NLP techniques for its detection. It covers:
  - A comprehensive overview of *Business Email Compromise*, including its definition, mechanisms, and significance in modern cyber threat landscapes.
  - *Statistics and trends*, including financial losses and attack frequency from IC3 and industry reports.
  - A breakdown of the *anatomy, methods, and strategies* used in BEC attacks, such as account takeover, invoice fraud, and executive impersonation.
  - An introduction to *Authorship Verification (AV)*, covering stylometric features and contrasting traditional hand-engineered and modern deep learning-based AV techniques.
  - A conceptual overview of *NLP* and *Transformer-based models*, focusing on their evolution, architecture, and role in modelling linguistic deception.
  - A review of *Multi-Task Learning (MTL)* principles, highlighting its advantages, relevance to NLP, and suitability for jointly tackling BEC detection and authorship verification within a unified framework.
- **Chapter 3: Literature Review** Critically surveys the academic and industry landscape surrounding (BEC) detection and prevention. It includes:

  - A structured comparison of technical (e.g., rule-based, ML, NLP, cryptographic) and non-technical (e.g., awareness training, policy) countermeasures adopted to mitigate BEC threats.
  - A comprehensive synthesis of datasets used in BEC research, including public corpora (Enron, TREC), proprietary datasets (e.g., BEC-Guard), and simulated multilingual datasets, highlighting limitations in coverage and realism.
  - An in-depth performance comparison across diverse BEC detection methods, with attention to metrics like accuracy, precision, and false positive rate.
  - Identification of three persistent gaps in the literature.
  - A cross-reference to thesis chapters that directly address each gap through taxonomy creation, content-based detection, authorship verification, and a unified NLP-based MTL framework.

- **Chapter 4: Methodology** Outlines the research design, methodological choices, and experimental processes that underpin this thesis. It includes:
  - A mixed-methods strategy that combines quantitative experimentation with qualitative thematic analysis to achieve methodological triangulation and ensure research validity.
  - A phase-wise progression—from the systematic literature review and taxonomy development to model construction, evaluation, and final integration.
  - Explicit alignment of research questions, methodological phases, and outcomes, assessed with clearly defined metrics.

- **Chapter 5: BEC Taxonomy** Proposes a five-axis taxonomy to address the lack of structured classification schemes in (BEC) research. The taxonomy systematically categorises BEC incidents along five dimensions: attack anatomy, adversary methodology, target roles, countermeasures, and detection challenges. It provides:
  - A detailed framework for analysing and comparing BEC incidents, enabling more consistent threat modelling and defence design.
  - Illustrative real-world case studies—including Treasure Island and an insurance broker firm—to validate the taxonomy's descriptive coverage and applicability.
  - A bridge between conceptual classification and technical design, setting the foundation for the content and authorship detection models introduced in later chapters.

- **Chapter 6: Synthetic Dataset Creation** Addresses the lack of publicly available datasets for Business Email Compromise (BEC) and authorship verification by introducing a purpose-built synthetic corpus. This chapter includes:
  - A structured nine-stage generation pipeline involving real BEC seed cases, prompt engineering, LLaMA-based text generation, and quality control.
  - Subsets tailored for both semantic deception and stylistic mimicry, including synthetic BEC attacks, authorship mimicry, and impersonation-based emails.

  - Integration of phishing corpora and validation using BLEU/ROUGE metrics and human annotation to ensure linguistic realism and adversarial plausibility.
  - Ethical and legal safeguards to ensure research compliance and responsible data use.

- **Chapter 7: Transformer-Based BEC Detection Model** Presents the first experimental contribution—a deep learning model for detecting BEC using email body content. This chapter includes:
  - A hybrid architecture combining contextual embeddings with BiLSTM for sequential modeling.
  - A structured experiment comparing this model to classical baselines (TF-IDF with logistic regression, Random Forest, and XGBoost).
  - Evaluation on two benchmark datasets (Fraud, TREC07), demonstrating state-of-the-art performance across precision, recall, F1-score, and accuracy.
  - A targeted "mimic" test using AI-generated emails that imitate trusted senders' styles, revealing that content-only models struggle when deception mimics genuine writing.
  - Motivation for stylistic authorship verification as a necessary complement to semantic detection.

- **Chapter 8: Siamese Network for AV** Introduces the second technical contribution: a transformer-based Siamese model designed to verify authorship in business emails as a defence against stylistic impersonation. This chapter includes:
  - Justification for applying Siamese networks in the AV task, focusing on writing-style consistency.
  - The architecture of BiBERT-AV, which combines BERT embeddings with BiLSTM layers in a pairwise contrastive framework.
  - Empirical results across varying author pool sizes (2 to 50) and evaluation on a synthetic dataset of LLM-generated mimic emails.
  - Comparative analysis showing that BiBERT-AV significantly outperforms traditional and transformer-only AV models.
  - Discussion of AV's operational role in BEC defence, especially for detecting high-fidelity impersonation.

- **Chapter 9: Multi-Task Learning (MTL) Framework for BEC and AV** Presents the final technical contribution—a unified framework that jointly performs (BEC) detection and AV through Multi-Task Learning. This chapter includes:
  - Design motivation for integrating semantic (BEC) and stylistic AV analysis using a shared encoder with task-specific heads.
  - Comparative evaluation against sequential transfer learning baselines, showing improvements in accuracy, F1-score, and false-positive reduction.

  - Analysis of model robustness, generalization, and efficiency through ablation studies, ROC curves, and inference time benchmarks.
  - Discussion of real vs. synthetic data generation, author overlap constraints, and embedding visualizations that validate cross-task feature learning.

- **Chapter 10: Conclusions and Future Work** Synthesizes the thesis contributions, discusses key limitations, and proposes future research directions. This chapter includes:
  - A structured review of how each research question was addressed, supported by empirical evidence across chapters.
  - Critical reflections on thematic limitations, including dataset realism, adversarial robustness, multilingual constraints, and deployment scalability.
  - Future research paths involving multilingual and domain-adaptive models, explainable AI for BEC detection, and psycholinguistic signals for authorship verification.

Chapter 2

Background

This chapter provides an overview of Business Email Compromise (BEC) fraud. It begins by defining BEC and discussing its significance in today's cyber threat landscape. Key financial statistics and trends are then reviewed to highlight its growing impact. AV is introduced as a complementary approach for detecting stylistic inconsistencies in impersonation-based attacks.

The chapter then presents core concepts in NLP, with a focus on Transformer-based models such as BERT, as well as BiLSTM and Siamese networks for text representation and similarity. Finally, it introduces the paradigm of Multi-Task Learning (MTL), which enables the joint modelling of related tasks, such as BEC detection and authorship verification.

2.1 Business Email Compromise

Business Email Compromise (BEC) is a targeted form of phishing and email fraud that specifically exploits employees with access to sensitive or financial information. These attacks rely on impersonation and social engineering to deceive recipients into performing unauthorized actions, such as:(i) initiating financial transfers (e.g., fraudulent invoices), (ii) disclosing confidential data (e.g., employee records), and (iii) complying with requests from impersonated authority figures (e.g., executives or legal counsel).

The primary attack vector in BEC is linguistic and contextual rather than technical. Adversaries use reconnaissance to craft well-written, context-aware emails that emulate the tone, style, and behavioural patterns of known personnel. For example, executive impersonation attacks (also known as CEO fraud) use authority framing, urgency, and role-specific phrasing to socially engineer finance teams into fast-tracking payments. Similarly, vendor fraud attacks involve the hijacking or spoofing of supplier communications to redirect invoice payments Federal Bureau of Investigation (2024). These tactics avoid triggering conventional threat detection systems, which typically scan for known malware signatures or obvious anomalies in email headers.

Although BEC and phishing are both categorized as social engineering attacks, they differ significantly in their attack strategies and detection challenges. Conventional phishing typically involves bulk-distributed emails containing broadly targeted lures, such as fabricated account alerts or password reset requests, and often includes detectable technical indicators like spoofed URLs or malicious attachments. In contrast, BEC attacks are highly targeted, linguistically sophisticated, and context-specific, crafted to impersonate internal stakeholders and align with legitimate business communication. Framing BEC merely as a subcategory of phishing overlooks its unique reliance on semantic manipulation and identity deception.

BEC fraud typically progresses through a series of well-defined stages:

- **Reconnaissance:** Attackers begin by gathering detailed information about the target organization—such as organizational charts, email communication patterns, and key personnel details—to tailor their approach. This stage is well-described by Saud Al-Musib et al. (2021), who emphasize the role of intelligence gathering in shaping effective BEC strategies.
- **Initial Compromise:** Using the acquired intelligence, attackers gain initial access by either compromising a legitimate email account or establishing a fraudulent relationship with a trusted individual. This step is commonly observed in FBI Public Service Announcements (PSAs) on BEC incidents (Service-Announcement, 2024).
- **Infiltration:** With access secured, the attacker monitors internal communications to determine the optimal time to launch a financial fraud attempt. As reported in the IC3 Elder Fraud Report (Federal Bureau of Investigation, Internet Crime Complaint Center (IC3), 2024), such monitoring often continues for days or weeks to ensure credibility and timing.
- **Execution:** An urgent, deceptive request—often impersonating a high-ranking executive—is then sent to initiate an unauthorized transaction. Security (2017) describe how attackers often cite unavailability due to travel or meetings to discourage verification.
- **Exfiltration:** Once the funds are transferred, the attacker quickly moves the money to intermediary accounts, making recovery extremely difficult. The IC3 report for 2023 highlights how these tactics complicate financial tracking and law enforcement efforts (FBI Internet Crime Complaint Center, 2024).

Understanding the operational stages and linguistic sophistication of BEC attacks provides a foundation for assessing their real-world impact. Over the past decade, BEC has evolved from isolated incidents into a global threat with substantial financial consequences. To contextualize its growing prominence within the broader cybercrime landscape, the following section presents statistical insights and trend analyses from major cybersecurity and law enforcement reports.

2.1.1 Statistics and Trends

According to FBI Internet Crime Complaint Center (2024), the FBI's Internet Crime Complaint Center (IC3) reported adjusted losses from BEC fraud reaching $2.94 billion in 2023, based on 21,489 complaints—continuing an upward trajectory from $2.74 billion in 2022 and $2.39 billion in 2021.

More recent data from the IC3's 2024 report suggest that BEC remains one of the most financially damaging forms of cyber-enabled fraud. Although losses in 2024 slightly decreased to approximately $2.77 billion across 21,442 cases, BEC still ranked second only to investment fraud in terms of total financial impact (FBI Internet Crime Complaint Center (IC3), 2025).

Viewed across the three-year period from 2022 to 2024, BEC accounted for some of the highest cumulative losses—aggregating to nearly $8.5 billion (NACHA, 2024).

These figures underscore BEC's enduring severity. Despite year-to-year fluctuations, the multi-year trend remains alarmingly high, reinforcing the need for specialized, content-based defenses capable of recognizing impersonation and deception tactics absent in traditional metadata-based systems.

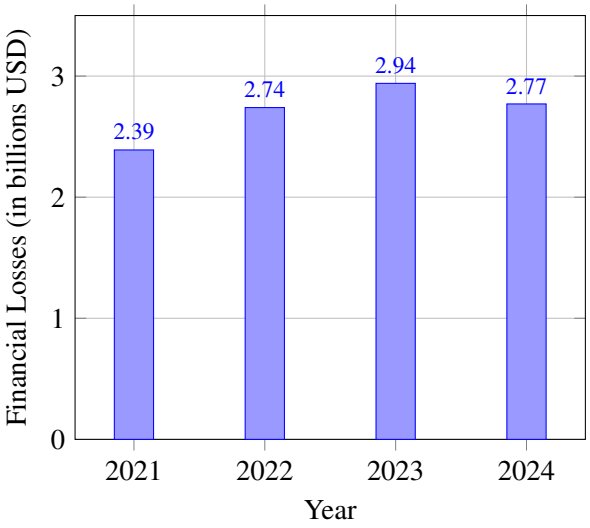


FIGURE 2.1: Financial losses due to BEC from 2021 to 2024, based on FBI IC3 data.

Further reinforcing the economic impact of BEC is *Figure 2.1*, which shows the ranking of cybercrime categories by total complaint losses in 2023. BEC remains one of the most financially devastating types of cybercrime, as seen in its prominent position within the top five categories. This emphasizes the pervasive threat that BEC poses, particularly in the context of broader cybersecurity challenges.

Moreover, *Table 2.1* provides an overview of the key research objectives in BEC detection, highlighting the critical issues being addressed in the literature. These include the deceptive

techniques employed by attackers, strategies for impersonation, and the use of linguistic manipulation, which are central to understanding the dynamics of BEC attacks. The diversity of research objectives indicates the complexity of BEC and the need for multifaceted detection approaches that address both the technical and psychological components of these attacks.

TABLE 2.1: Summary of BEC Fraud Objectives

Source	Description	Objectives
Zweighaft (2017)	The attacker poses as a lawyer or representative of the law firm supposedly in charge of the company's legal matters and requests confidential information.	Stealing confidential, private information such as financial records, legal documents, and intellectual property. **Example**: An attacker impersonates the company's legal advisor and requests copies of recent merger and acquisition documents.
King (2019)	The attacker uses a hacked executive's or employee's email account to make requests that appear legitimate to internal staff.	Financial or confidential information requests that appear to come from within the company, aimed at unauthorized fund transfers or data breaches. **Example**: An attacker uses a compromised CFO's email to instruct the finance department to change the bank account details for the next payroll run.
Cross and Gillett (2020)	Corporate fraud involving the identity theft of a senior member of an organization. The attacker sends emails asking for urgent financial transactions or access to confidential documents.	Urgent financial or confidential information requests aimed at diverting company funds or gaining access to sensitive information. **Example**: An email appearing to be from the CEO urgently requests the transfer of $100,000 to a new supplier's account.
Spangler (2021)	Detailed the BEC method and strategies employed by attackers to deceive targets into disclosing critical information.	Educating organizations on the various tactics used in BEC scams, helping them develop better preventive measures and response strategies. **Example**: Training sessions simulate BEC scenarios to help employees recognize and respond to suspicious emails effectively.

Business Email Compromise (BEC) continues to rank among the most financially damaging and operationally sophisticated forms of cybercrime. Reports from Microsoft, IBM, and the UK's National Cyber Security Centre (NCSC) consistently identify BEC as a top-tier threat due to its reliance on targeted deception rather than technical exploits. Despite its increasing prevalence and financial impact, BEC remains under-represented in academic research—particularly in the domains of machine learning and NLP. This gap highlights the urgent need for advanced, content-based detection approaches capable of capturing the subtle linguistic and behavioural cues that characterize BEC attacks Atlam and Oluwatimilehin (2022).

2.2 Authorship Verification (AV)

Authorship verification (AV) is considered one of the three primary domains of Automatic Authorship Identification (AAI)—alongside authorship attribution and authorship identification—as described by Brocardo et al. (2013). The AV task involves determining whether a new digital text was authored by a specific individual when a candidate author is presented with a set of known texts. Typically, this is framed as a binary classification problem, as depicted in Figure 2.2.

The primary goal of AV is to identify writing style consistencies and variations to verify the authorship of a given text. This process has numerous applications, including detecting plagiarism, identifying anonymous authors, and forensic document analysis. Additionally, AV plays a critical

role in social media forensics by uncovering aliased accounts and in information security by enabling continuous user authentication.

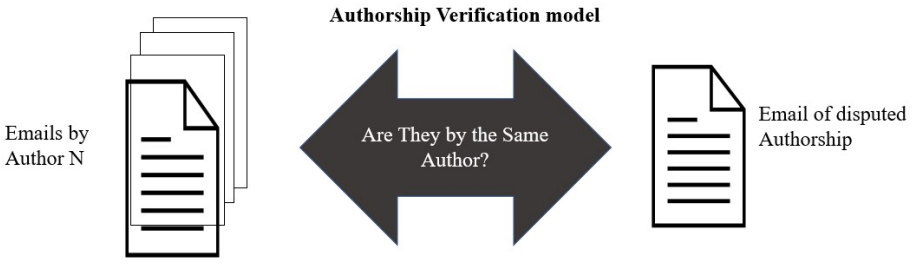


FIGURE 2.2: An authorship verification problem

2.2.1 Stylometric Features in Writing

Stylometric analysis forms the foundation of AV by examining linguistic features inherent in an individual's writing style Stamatatos (2009). These features can be broadly categorized into:

- Lexical Features: Word usage, frequency, average word length, and vocabulary richness.
- Syntactic Features: Sentence structure, punctuation patterns, and grammar usage.
- Structural Features: Document layout, paragraph organization, and formatting preferences.
- Semantic Features: Word semantics and topic modeling to analyze the context and meaning.

These stylometric features offer clues into the author's unique linguistic patterns, forming a distinct "signature" that can be used for verification.

2.2.2 Traditional vs. Modern AV Techniques

The development of AV methods can be broadly divided into two conceptual paradigms: traditional stylometric approaches and modern representation learning frameworks.

Traditional techniques rely on manually engineered features that reflect an author's writing style. These include lexical patterns, syntactic structures, and statistical distributions of character or word usage. Common methods involve modeling stylistic fingerprints using character *n*-gram profiles or computing stylistic dissimilarity through distance-based metrics such as the out-of-place measure proposed by Kešelj et al. (2003). Other classical approaches include the *unmasking* technique using support vector machines (SVMs) introduced by Koppel and Schler (2004), the profile-based dissimilarity approach that achieved notable success in the PAN-AV'14 competition Potha and Stamatatos (2014), and compression-based similarity kernels Halvani et al. (2017). While these models are generally fast, interpretable, and effective for short texts, their reliance on surface-level features often limits their robustness under domain shifts, cross-topic variation, or intentional obfuscation by adversaries.

Modern techniques, by contrast, treat authorship verification as a representation learning problem. These methods aim to capture deeper semantic and syntactic cues by learning task-specific embeddings that generalize across contexts. Siamese and triplet neural architectures are commonly used to project text pairs into a shared embedding space, where same-author texts are positioned closer together than texts by different authors. For example, convolutional neural networks (CNNs) applied to character-level *n*-grams have been used to construct pairwise similarity models Araujo-Pino et al. (2020). More recently, fine-tuned transformer-based language models have become the dominant paradigm. Models such as BertAA leverage the bidirectional contextual representations of BERT to learn stylistic patterns beyond handcrafted features Fabien et al. (2020). Variants like Longformer are designed to handle long documents efficiently, enabling analysis of emails and other extended texts Ordoñez (2020). Siamese BERT architectures have also been applied to authorship verification in email domains Tyo et al. (2021), while chunked encoding strategies have proven effective for low-resource or short-form datasets Peng (2021). These transformer-based approaches benefit from self-attention mechanisms that capture both local and global dependencies, making them particularly resilient to style variation and adversarial manipulation.

Overall, the transition from feature-based to embedding-based methods reflects a shift toward more expressive, generalizable models capable of handling the complexity of real-world authorship verification tasks.

2.3 Natural Language Processing (NLP) and Transformer Models

NLP is a branch of artificial intelligence (AI) focused on enabling machines to understand, interpret, and generate human language Chowdhury (2003). Over time, NLP has evolved from rule-based approaches to modern deep learning methods, with Transformer-based architectures revolutionizing the field. This section provides an overview of key NLP advancements, particularly Transformer models, and their relevance to the research.

2.3.1 Evolution of NLP

The development of NLP has progressed through three major methodological phases, each reflecting advances in both computational capabilities and linguistic modeling:

- **Rule-Based Methods:** Early NLP systems were built on manually crafted rules and deterministic grammars to encode syntactic and semantic knowledge. This approach was famously introduced by Chomsky (1957), whose work laid the foundation for formal language theory. While interpretable and effective for constrained tasks, these systems lacked robustness to linguistic variability and ambiguity.
- **Statistical and Traditional Machine Learning Approaches:** The introduction of probabilistic models—such as n-gram language models—marked a shift toward data-driven NLP. Brown et al. (1990) demonstrated how statistical techniques could model language regularities at scale. These were later extended using classical machine learning algorithms (e.g., SVMs, CRFs), which allowed the modeling of more complex structures but still required extensive feature engineering.
- **Neural and Deep Learning-Based Models:** The emergence of deep learning transformed NLP by enabling end-to-end learning from raw text. Architectures such as recurrent neural networks (RNNs), long short-term memory (LSTM) networks, and transformer-based models captured richer contextual dependencies. In particular, Vaswani et al. (2017) introduced the Transformer architecture, which significantly advanced performance on tasks like machine translation, sentiment analysis, and text classification.

2.3.2 Introduction to Transformer Models

Transformer architectures, first introduced by Vaswani et al. (2017), represent a fundamental shift in neural sequence modelling by eliminating the need for recurrence. Instead, Transformers rely entirely on self-attention mechanisms, enabling more efficient and scalable modelling of long-range dependencies in text.

The core components of the Transformer architecture include:

- **Self-Attention Mechanism:** Allows the model to dynamically compute pairwise interactions between all tokens in a sequence, capturing context-sensitive representations that reflect both local and global dependencies.
- **Encoder–Decoder Structure:** In the original formulation, the encoder maps input tokens to contextual embeddings, while the decoder generates output tokens autoregressively, attending to both previous outputs and encoder states.

- **Parallelization and Scalability:** Unlike recurrent models (e.g., RNNs or LSTMs), Transformers enable parallel computation across input tokens, significantly improving training efficiency and enabling scaling to very large datasets.

As shown in Figure 2.3, this architecture forms the foundation of modern pretrained language models, many of which have become the de facto standard for downstream NLP tasks.

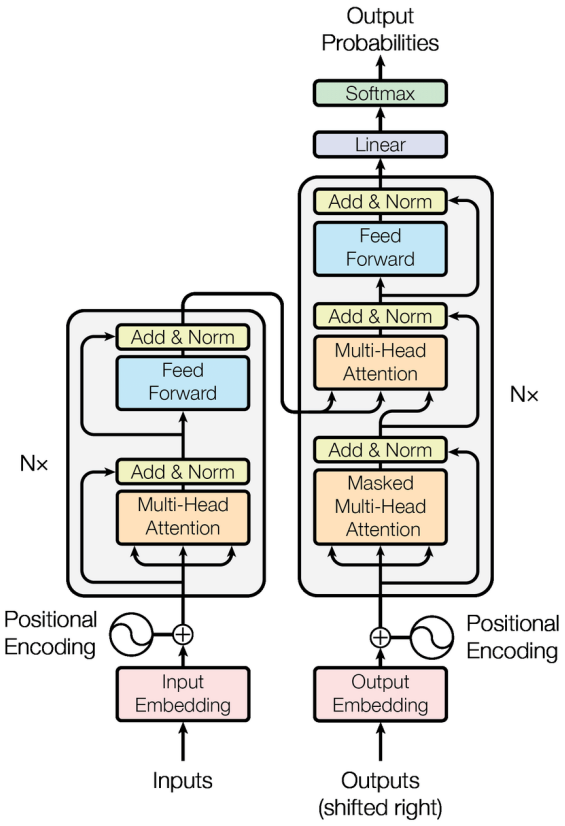


FIGURE 2.3: The Transformer model architecture as introduced by Vaswani et al. (2017).

2.3.3 Pre-trained Transformer Models in NLP

Pre-trained Transformer-based language models have become the foundation for most state-of-the-art NLP systems. These models are initially trained on large-scale text corpora using self-supervised learning objectives, and then fine-tuned on specific downstream tasks such as classification, question answering, or authorship verification.

- **BERT (Bidirectional Encoder Representations from Transformers):** Introduced by Devlin et al. (2018), BERT leverages masked language modelling (MLM) and next sentence prediction (NSP) to learn deep bidirectional contextual representations. It has demonstrated strong performance across diverse NLP benchmarks, including GLUE and SQuAD, and serves as a foundational model for many subsequent variants.

- **DistilBERT:** Proposed by Sanh et al. (2019), DistilBERT is a compressed version of BERT obtained through knowledge distillation. It retains most of BERT's representational power while significantly reducing model size and inference time, making it suitable for resource-constrained or real-time applications.
- **Other BERT-based Variants:** Several models extend or refine BERT's architecture and training objectives to enhance efficiency or performance:
  - **RoBERTa** removes the NSP objective and trains with more data and longer sequences;
  - **ALBERT** shares parameters across layers to reduce memory consumption;
  - **XLNet** introduces permutation-based pretraining to capture bidirectional context without masking.

These pre-trained models are commonly fine-tuned on task-specific datasets for applications such as sequence classification, question answering, and authorship verification.

2.3.4 Long Short-Term Memory (LSTM) and Bidirectional LSTM (BiLSTM)

Recurrent Neural Networks (RNNs) are designed for sequence modeling tasks but suffer from vanishing gradient issues when capturing long-range dependencies. Long Short-Term Memory (LSTM) networks, introduced by Hochreiter and Schmidhuber (1997), address this limitation by incorporating gated mechanisms that regulate information flow across time steps. An LSTM cell includes three gates—the input gate, forget gate, and output gate—which jointly control what information to retain, discard, or output at each step.

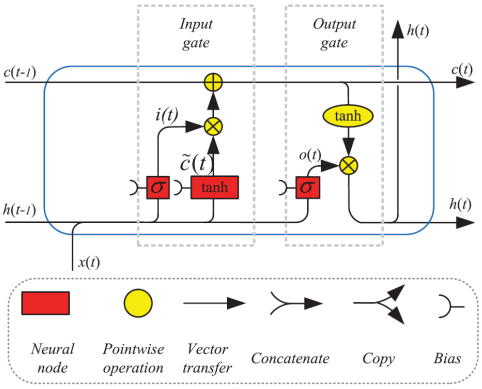


FIGURE 2.4: Structure of an LSTM cell. Adapted from Yu et al. (2019).

LSTMs are widely used in NLP tasks such as sentiment analysis, language modeling, and text classification due to their ability to maintain context over long sequences.

Bidirectional LSTM (BiLSTM) extends the standard LSTM by processing the input sequence in both forward and backward directions. This dual pass enables the model to capture dependencies

from past and future contexts simultaneously. The outputs from both directions are typically concatenated to form a richer representation of each token.

2.3.5 Siamese Networks for Text Similarity

Siamese networks are a class of deep learning architectures designed to determine the similarity between two inputs by learning a shared representation space. Initially introduced for tasks like signature verification Bromley et al. (1993), they have become widely used in NLP for comparing text pairs. Figure 2.5 shows the structure of a Siamese network.

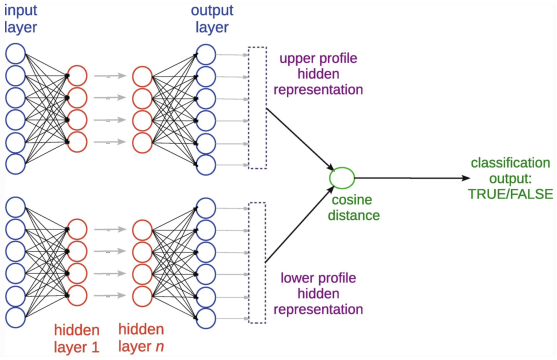


FIGURE 2.5: Representation of the Siamese neural network model. Cosine distance measures the similarity between input pairs as the final output Chicco (2021).

Key features of Siamese networks include:

- Shared Weights: Two identical subnetworks process inputs, ensuring consistent representation learning.
- Similarity Metrics: Outputs are compared using metrics like cosine similarity or Jaccard similarity to determine how closely related the inputs are.

2.4 Multi-Task Learning (MTL)

Multi-Task Learning (MTL) is a machine learning paradigm in which a single model is trained to solve multiple tasks concurrently, rather than optimizing each task independently. The approach was originally formalized by Caruana (1997), who demonstrated that task relatedness can be exploited by enabling shared representations. This allows the model to leverage common linguistic or structural patterns across tasks. Such shared inductive bias improves generalization, particularly in settings with limited labeled data, as emphasized by Ruder (2017).

Unlike single-task learning (STL), which focuses on a single objective, MTL introduces regularization through parameter sharing. When tasks are semantically aligned or exhibit similar input structures, this joint training facilitates the learning of more robust and transferable representations.

These advantages make MTL particularly effective in NLP, where many tasks share linguistic patterns or semantic structures.

For instance, MTL has been successfully applied to sentiment analysis and topic classification (Sebbaq et al., 2023), sequence labeling tasks such as named entity recognition (NER) and part-of-speech tagging (POS) (Yang and Shang, 2019; Zuo and Zhang, 2020), as well as question answering and multilingual translation (Dou et al., 2024; Wang et al., 2017; Xiao et al., 2022). These studies highlight the value of MTL in enhancing both efficiency and generalization across diverse language understanding tasks.

2.4.1 Benefits of Multi-Task Learning

MTL offers several advantages that are particularly valuable in NLP and security-sensitive applications:

- **Parameter Efficiency:** A unified architecture reduces model redundancy by sharing layers across tasks, minimizing training and deployment overhead.
- **Improved Generalization:** Auxiliary tasks act as inductive regularizers, reducing overfitting and encouraging the model to learn features that generalize well across related objectives. This principle was first demonstrated by Caruana (1997), who showed how joint training improves generalization by capturing task-invariant patterns.
- **Effective Use of Limited Data:** MTL facilitates knowledge transfer from high-resource to low-resource tasks through shared representations, making it ideal for domains where annotated data is scarce. As highlighted by Ruder (2017), MTL is particularly advantageous when data sparsity would otherwise limit single-task performance.
- **Cross-Task Synergy:** When tasks are complementary—such as classification and verification—their joint optimization can lead to mutual performance gains through shared supervision.

These advantages make MTL particularly effective in NLP, where many tasks share linguistic patterns or semantic structures.

2.5 Chapter Summary

This chapter provides an overview of the foundational concepts supporting this research. It examines Business Email Compromise (BEC) attacks, focusing on their impacts, stages, and strategies such as account seizure and impersonation. It also explores AV, discussing stylometric features, traditional versus modern techniques, and applications in enhancing email security.

NLP and Transformer-based models are introduced, highlighting their evolution, core mechanisms, and pre-trained models like BERT and DistilBERT. The relevance of Siamese networks for text similarity tasks, particularly in AV, is also covered.

Finally, MTL is discussed, emphasizing its advantages, applications in NLP, and its role in integrating BEC detection and AV. Together, these topics establish the foundation for the proposed methodologies in later chapters.

Chapter 3

Literature Review

3.1 Introduction

The increasing reliance on digital communication has significantly reshaped organizational workflows, particularly in finance and enterprise environments. While the use of email has enhanced operational efficiency and enabled rapid transactions, it has also created new opportunities for exploitation. Among the most financially damaging cyber threats is Business Email Compromise (BEC), a form of social engineering in which attackers impersonate trusted individuals—such as executives, vendors, or clients—to deceive victims into transferring funds or disclosing sensitive information.

The primary aim of this chapter is to establish a comprehensive understanding of the current state of Business Email Compromise (BEC) detection research in order to address *SRQ1: What approaches currently exist for detecting BEC attacks, and what are their respective strengths and limitations?* To address this question, the current state of research on BEC detection is critically reviewed. This includes both technical and non-technical countermeasures, such as rule-based filters, metadata analysis, content-aware models, and behavioural profiling.

The chapter also evaluates the datasets used in BEC research. Finally, concludes by identifying open challenges in current approaches. These limitations motivate the architectural decisions and methodological contributions presented in subsequent chapters of this thesis.

[0]This chapter is based on the published article: Almutairi, A. M., Kang, B., & Al Hashimy, N. (2025). *Business email compromise: A systematic review of understanding, detection, and challenges*. *Computers & Security*. doi:10.1016/j.cose.2025.104630.

3.2 BEC: Literature Review

3.2.1 BEC: Systematic Literature Review Methodology

This literature review adopts a systematic literature review (SLR) approach to provide a comprehensive and rigorous synthesis of Business Email Compromise (BEC) research. The review follows the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) framework and established guidelines for systematic reviews in cybersecurity research Kitchenham and Charters (2007). This ensures methodological transparency, replicability, and consistency in how the literature was identified, selected, and analyzed.

The methodology, illustrated in Figure 3.1, documents each stage of the process.

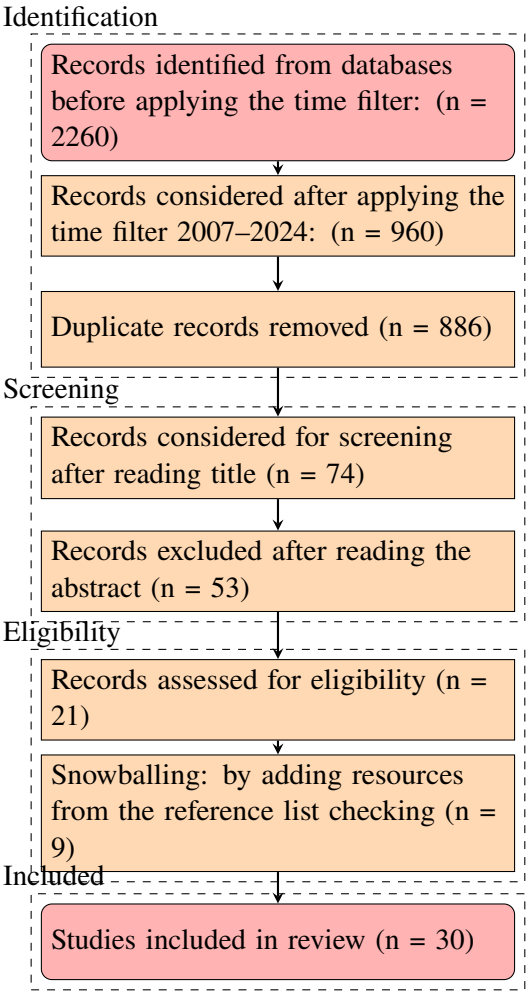


FIGURE 3.1: PRISMA workflow for study selection in this SLR.

3.2.1.1 Search Strategy

A Boolean search query was applied across databases such as IEEE Xplore, ACM Digital Library, Scopus, Web of Science, SpringerLink, and ScienceDirect. Keywords combined core concepts of

Business Email Compromise with detection and prevention techniques:

```
("Business Email Compromise" OR "BEC" OR "CEO fraud" OR "email fraud")
AND ("detection" OR "prevention" OR "machine learning" OR "NLP"
OR "cybersecurity")
```

3.2.1.2 Screening and Eligibility

The four PRISMA stages were:

- **Identification:** 2,260 records initially retrieved.
- **Screening:** 74 studies shortlisted after title and abstract review.
- **Eligibility:** 21 studies met full-text assessment criteria.
- **Inclusion:** Final corpus of 30 peer-reviewed studies, including 9 added through snowballing.

3.2.1.3 Quality Assessment

All included studies were evaluated using five criteria: research clarity, dataset transparency, empirical validation, methodological rigor, and relevance to BEC. Inter-rater agreement on a random subset achieved Cohen's $\kappa = 0.89$, confirming high consistency.

3.2.2 Countermeasures Against BEC Fraud

This section examines how companies and researchers have attempted to combat BEC fraud by proposing and evaluating a range of countermeasures. Specifically, presents a comprehensive classification of BEC detection and prevention techniques—both technical and non-technical—drawn from the studies surveyed.

Guided by the well-known People–Process–Technology (PPT) triad in security research, we define *technical* controls as technology-centric solutions (e.g., rule-based filters, ML/NLP models, cryptographic schemes) and *non-technical* controls as people- and process-centric measures (e.g., training, human verification, governance policies). This socio-technical framing moves beyond an intuitive split and offers a structured lens for comparing robustness, scalability, and deployment realism across studies.

3.2.2.1 Technical Countermeasures

Various technical mitigation proposals have been discussed in the literature. These countermeasures can be broadly divided into two main categories: *Traditional Rule-based Methods* and *Machine Learning-based Solutions*.

Furthermore, Table 3.1 summarizes the main technical detection techniques, comparing their reported results and highlighting key findings from recent BEC studies.

Traditional Rule-based Methods Scholars and industry experts have collaborated extensively to develop software defences and risk mitigation techniques that enterprises can deploy to counter the rising threat of BEC fraud. As discussed by Meyers (2018), protective measures such as maintaining up-to-date software, enforcing end-point security, deploying anti-malware systems, and utilizing digital signatures for emails can reduce exposure to BEC threats.

Another effective approach involves analysing historical email patterns to detect anomalies in communication behaviour. For instance, Cidon et al. (2019) developed BEC-Guard, a system that applies statistical profiling of user behavior to flag suspicious emails and prevent fraud in enterprise environments.

Typosquatting, a tactic in which attackers register domain names that closely resemble legitimate ones, poses an additional threat that can, in some cases, be mitigated through proactive domain monitoring and early warning systems Mansfield-Devine (2016). Organizations can also employ blocklists and allowlists to prevent fraudulent email interactions. Blocklists restrict access from known malicious sources, such as compromised IP addresses and suspicious email domains Siadati et al. (2020): "If the recipient's email address, IP address, or another characteristic has been blacklisted, the session will be canceled before the email is received". Conversely, allowlists define trusted email senders, reducing false positives. A well-balanced strategy incorporating both blocklists and allowlists is crucial to ensuring seamless, legitimate communication while filtering out fraudulent messages effectively.

Statistical Methods Shahrivari et al. (2020) employed the Delphi technique, collecting feedback from thirty cybersecurity experts to validate BEC detection criteria. Their study highlighted that global financial losses from BEC fraud exceeded $26 billion and identified four key factors crucial for effective detection: recognizing email authenticity, detecting malicious mobile applications, identifying indicators of mobile malware, and discerning phishing attempts. Their approach, which combined expert-driven insights with statistical validation, achieved an accuracy rate of 92.5%.

Acar et al. (2019) conducted a large-scale analysis of malware attacks collected from two organizations between 2017 and early 2018, focusing on threat vectors, time series analysis, vulnerabilities, and social engineering tactics. Unlike earlier malware research, their study

TABLE 3.1: Comprehensive summary of BEC detection techniques and reported performance. A dash (–) indicates that the metric was not reported in the original paper.

Source	Method	Description	Contributions	Limitations	Dataset	Acc.	Prec.	Recall	F1	FPR
Maleki (2019)	ML	K-means on keywords + sentiment scores	Detects BEC; sender-side blocking	Lacks dynamic feature selection; scale issues	Enron	92%	–	–	93%	–
Cidon et al. (2019)	ML	Two-stage (metadata + content)	Real-time detection; industry adoption	Requires continual re-training	BEC-Guard	–	98.2%	96.9%	–	1.9×10^{-5}
Regina et al. (2020)	NLP	BERT + translation augmentation	Improves text-classification robustness	Small simulated corpus	Simulated	87%	–	–	–	–
Xiao and Jiang (2020)	ML	KNN and Bi-LSTM for spam/phish	High phishing accuracy	Needs larger corpus	TREC	95.3/%	91.7%	91.49%	91.58%	1.22%
Kurematsu et al. (2019)	ML	Author identification from e-mail text	ML-based author-ID approach	Uses only first 100 words	Enron	84%	–	–	–	–
Teerakanok et al. (2020)	Checksum	Invoice checksum validation	Simple; counters bogus invoices	Needs secure key; manual checks	Simulated	N/A	–	–	–	–
Vorobeva et al. (2021)	ML	TF-IDF + k-Means + LDA	Works on Russian & English corpora	No insider-attack coverage	2 308 mails (English/Russa)	95% / 75%	–	–	–	–
Almutairi et al. (2023)	ML	BERT + BiLSTM for BEC	Captures high-level linguistics	Resource-intensive	Fraud, TREC	99%	99.8 / 98.5%	99.7 / 98.6%	99.8 / 98.6%	–
Wickline (2021)	Whitelist/ML	Endpoint monitoring + app whitelisting	Blocks malicious files	High resource use; FP overhead	Enterprise	N/A	–	–	–	–
Brabec et al. (2023)	ML	CAPE – ensemble ML system	Covers multiple e-mail types	Limited data; explainability gap	Mixed organisations	–	>80%	–	–	–
Papathanasiou et al. (2024)	Cryptography	QR-code + MAC secure-mail scheme	Resists replay attacks	Key distribution; timestamp drift	Simulated	–	High	–	–	–
Sahoo and Rajitha (2019)	IDS	Fake vs real e-mail discrimination	High phishing accuracy	Requires representative corpus	Enron	98%	–	–	–	–
Siadati (2019)	IDS	Social-engineering deterrence system	Recall > 80%	Needs attacker-comm access	77 scammer accounts	–	–	>80%	–	0.3%
Haddon (2020)	Firewall	Analysis of exfiltration vectors	Real-world insights	Resource-heavy deployment	Case studies	N/A	–	–	–	–

concentrated on modern industrial malware samples. A key finding of their time-based analysis revealed that 93% of malware samples were distributed during weekdays, underscoring the targeted nature of these attacks and the influence of corporate email traffic patterns.

Checksum Approach Teerakanok et al. (2020) proposed a semi-automated method for verifying the authenticity and integrity of financial transactions using a checksum generated from critical transaction details. The process involves a secret delivery key issued by the invoice-issuing entity, which the supplier then uses to generate a checksum by combining essential transaction data and the secret key. If both parties produce an identical hash, the transaction is deemed legitimate. Their approach employs the SHA256 message-digest function and converts the hash to base 8 for added security.

Papathanasiou et al. (2024) introduced the BEC Defender algorithm, which employs cryptographic techniques such as Message Authentication Codes (MACs) and QR codes to verify the authenticity of email communications. The system utilizes Fernet encryption for secure data storage and SHA2 hashing to enhance the security of the registration process. While extensive testing across multiple email providers and operating systems demonstrated the algorithm's effectiveness, certain limitations remain. These include:

- Challenges in secure key distribution.
- A three-hour validation window, which, while adding security, may reduce usability.
- Potential inaccuracies in MAC address verification.
- A residual risk of replay attacks within the validation timeframe.

Despite these challenges, BEC Defender represents a promising cryptographic approach to mitigating email-based fraud.

Intrusion Detection System (IDS) Sahoo and Rajitha (2019) proposed an intrusion detection approach designed to distinguish between legitimate and fraudulent emails, thereby safeguarding users against phishing attacks and data breaches. Their method, applied to the Enron dataset, achieved a 98% accuracy rate.

Siadati (2019) focused on BEC attacks that impersonate coworkers, a category of social engineering threats that often bypass traditional phishing detection mechanisms due to their lack of common indicators such as malicious links or suspicious IP addresses. The study introduced a novel countermeasure aimed at disrupting attackers by monitoring their private communications and intercepting key resources (e.g., stolen passwords and fraudulent bank account details). Their system demonstrated a recall rate exceeding 80% and a false positive rate of 0.3%, highlighting its effectiveness in identifying impersonation attempts.

Whitelisting and Firewall Methods Haddon (2020) analyzed BEC attack vectors and data exfiltration risks, emphasizing network lockdowns, firewall restrictions, and up-to-date antivirus systems as key defense strategies. Their study provided real-world insights into evolving attack techniques and countermeasures. While these methods can enhance security, they require significant resources and may struggle to keep pace with rapidly evolving threats. Their evaluation was based on case studies and historical reports, without reporting a specific accuracy metric.

Opazo et al. (2017) proposed a client-side security mechanism that analyzes email headers for inconsistencies, logs alerts, and notifies enterprise administrators of potential threats. Their framework includes whitelisting trusted contacts, which reduces false positives while maintaining strict email security policies.

Wickline (2021) examined the effectiveness of modern antivirus solutions in detecting and mitigating malware threats. The study identified BEC, phishing, and spear phishing as primary attack vectors and highlighted how malware is leveraged to disrupt critical infrastructure and steal sensitive data. Additionally, the research noted that malware development surged during the COVID-19 pandemic, with 350,000 new malicious programs created daily, leading to a 40% increase in global malware volume.

Other Approaches While technical defences and detection models form the foundation of BEC mitigation, a number of studies have taken broader or more specialized perspectives to address complementary dimensions of the threat. These include organizational case studies, risk modelling frameworks, and legal or regulatory analyses. Together, these contributions enrich the understanding of BEC by highlighting its psychological, procedural, and institutional implications—extending beyond algorithmic detection and infrastructure-level controls.

Awah Buo (2020) examined the global rise of BEC fraud and presented a case study of Unatrac Holding Ltd. They conducted a detailed investigation into the psychological and sociotechnical impact of a successful BEC attack on both the organization and its employees.

Benaroch (2018) proposed a model modification approach for BEC risk management, where zero or more precautionary measures can be deployed in varying sequences. These measures have impulse-type effects to reduce uncertainty, and their impacts can be substitutive, complementary, or synergistic. This modelling approach enables both passive and proactive IT risk management.

Kolouch (2016) studied legal implications and potential criminal liabilities of phishing, scams, BEC, and other specialized cyberattacks. Their focus extended to international legal standards, including those defined in the Convention on Cybercrime, as well as the relevant laws within the Czech Republic.

Machine Learning-based Solutions Machine Learning (ML) has been widely and successfully applied to various business and research applications, including BEC detection.

Maleki (2019) proposed and tested a behavior-based detection model for compromised email accounts or machines. The model prevents fraudulent emails by blocking messages from compromised senders who fail to form a valid user profile from the recipient's perspective. Additionally, the system alerts legitimate account owners when a compromise is detected. Evaluated on the Enron Dataset, the framework achieved 92% accuracy and a 93% F1-score.

Cidon et al. (2019) introduced BEC-Guard, a two-stage detection system for identifying and blocking impersonation emails. The first stage analyzes email metadata (e.g., sender, receiver, CC, BCC fields) to detect anomalous patterns. If flagged, the email proceeds to content-based analysis, which employs NLP and link verification. The text classifier uses TF-IDF with unigrams and bigrams (10,000 features), while the link classifier flags small or newly created websites. The combined system reported 98.2% precision, 96.9% recall, and an extremely low false positive rate of 0.000019% (1 in 5,260,000 emails). Despite its success, continuous retraining is necessary to counter evolving attack strategies.

Kurematsu et al. (2019) developed an ML-based author identification model for BEC detection, focusing on writing style analysis. Unlike traditional spam filters, this approach relies on author profiling, analyzing the first 100 words of an email body. Evaluated on the Enron dataset, the system achieved 84% accuracy, highlighting its potential for authorship verification in email security.

Vorobeva et al. (2021) proposed a BEC detection method based on writing style analysis. Their feature set included word n-grams, three-gram phrases, day-of-week, time sent, message urgency, and email headers. Using Linear Support Vector Classification (LSVC) with feature scaling, their system achieved 95% accuracy for English emails and 75% accuracy for Russian emails.

Xiao and Jiang (2020) introduced a phishing and spam detection system using K-Nearest Neighbors (KNN) and Bi-LSTM. Their approach significantly reduced false positives while maintaining high accuracy. Their experiments on the TREC06P dataset resulted in 95.27% accuracy (KNN), 91.51% accuracy (Bi-LSTM), 91.75% precision, 91.49% recall, 91.58% F1-score, and a false positive rate of 1.22

Brabec et al. (2023) developed CAPE, a modular and adaptive BEC detection system designed for Security Operations Centers (SOC). CAPE integrates multiple ML models and applies a Bayesian framework for continuous refinement. Over two years, CAPE's precision remained consistently above 80%, demonstrating its reliability in real-world applications. However, its performance heavily depends on data availability, operational costs, and explainability.

NLP Methods Complementing the broader ML landscape, NLP techniques emphasize textual content and linguistic cues, which are especially relevant for deception detection in email communications. Regina et al. (2020) introduced a task-agnostic augmentation system that combines BERT, reverse translation, and heuristic-based NLP enhancements. Their method

achieved 96% balanced accuracy on a BEC detection task, demonstrating the value of language-specific augmentation techniques.

While machine learning and NLP-based approaches demonstrate significant effectiveness in detecting BEC through behavioural modelling, statistical profiling, and linguistic analysis, these technical solutions represent only part of the broader defence landscape. The literature also highlights a range of non-technical countermeasures that focus on organizational practices, procedural safeguards, and policy-level interventions. The following section reviews such approaches, emphasizing their role in strengthening institutional resilience against BEC threats in practical, real-world contexts.

3.2.2.2 Non-Technical Solutions

Alongside technical countermeasures, non-technical approaches are critical in mitigating BEC fraud. These methods focus on human factors, policies, and awareness to complement automated systems.

Awareness Training Employee education is a vital preventative tool against BEC fraud. Several studies such as: Mansfield-Devine (2016); Binks (2019); Ross (2018) have demonstrated that company-wide training—via workshops, phishing simulations, and role-playing exercises—enhances employees' abilities to recognize fraudulent emails. As noted by Nehme and George (2018), organizations must continually update and engage their staff to reinforce critical security behaviours.

Human Verification The FBI (2021) advises that users verify suspicious URLs, check for typographical errors in email addresses, and confirm the authenticity of requests through secondary channels. Human verification acts as a crucial backup when technical systems fail to flag sophisticated impersonation attacks.

Policies and Guidelines Robust organizational policies, such as multi-factor authentication and dual-approval workflows, are essential. Studies by Meyers (2018) and Burns et al. (2019) illustrate that governance frameworks—where high-value transactions require cross-checks—reduce the risk of fraudulent transfers. Additional research Susanti et al. (2023); Ogwo-Ude (2023) emphasizes the importance of integrating cybersecurity policies, risk management systems, and regulatory compliance (e.g., ISO 27001:2013) to further mitigate BEC threats.

Table 3.2 summarizes several non-technical solutions, including their strengths and limitations.

TABLE 3.2: Summary of Non-Technical Solutions for BEC Fraud Detection

Source	Method	Description	Strengths	Limitation
Mansfield-Devine (2016)	Awareness Training	Employee education on phishing and BEC fraud.	Enhances recognition skills; reduces susceptibility.	Requires continuous updates and engagement.
Binks (2019)	Awareness Training	Company-wide training to minimize phishing assaults.	Comprehensive awareness; effective simulation exercises.	Implementation may be resource-intensive.
Ross (2018)	Awareness Training	Simulated attack training to understand BEC indicators.	Improves response and recognition.	Needs regular updates to match evolving tactics.
Zweighaft (2017)	Awareness Training	BEC testing and training across organization levels.	Builds a proactive, skeptical culture.	Requires ongoing resource allocation.
Nehme and George (2018)	Awareness Training	Programs to educate employees on phishing, social engineering, and risks.	Enhances critical analysis and email verification skills.	Dependent on continuous engagement.
Lazarus (2024)	Awareness Training	Qualitative analysis of cybercriminal networks and tactics.	Provides insights into criminal methods.	Focuses on a single case study; limited generalizability.
Papathanasiou et al. (2023)	Awareness Training	Examines social structures of BEC criminals via interviews.	Offers insider perspectives on social engineering.	Limited by focus on a specific criminal group.
FBI (2021)	Human Verification	Advises users to verify URLs and sender details.	Simple, direct approach to authentication.	Relies heavily on user diligence.
Meyers (2018)	Policies and Guidelines	Recommends multiple sign-offs for significant transactions.	Adds verification layers; reduces unilateral risk.	May slow down legitimate processes.
Burns et al. (2019)	Policies and Guidelines	Suggests a business governance framework for high-value transactions.	Establishes formal procedures; deters fraudulent requests.	Implementation can be complex and time-consuming.
Susanti et al. (2023)	Policies and Guidelines	Emphasizes robust cybersecurity policies and training.	Enhances overall cybersecurity posture.	Does not provide direct technical defense.
Ogwo-Ude (2023)	Policies and Guidelines	Recommends advanced email authentication and incident response plans.	Offers comprehensive, coordinated protection.	Requires interdepartmental coordination.

3.2.3 Datasets Used in BEC Research

High-quality datasets are crucial for developing and evaluating BEC detection systems. However, due to privacy concerns and the sensitive nature of business communications, publicly available BEC datasets are scarce. Researchers rely on a combination of public datasets, proprietary corpora, and simulated data.

For example:

- **Enron Email:** A public dataset containing approximately 500,000 emails from 150 employees. It has been widely used in prior studies on organizational email behavior and security applications, including the works of Maleki (2019), and Kurematsu et al. (2019).
- **TREC:** A public dataset with about 50,000 emails (35,000 spam and 15,000 non-spam messages) used for benchmarking spam and ham classification methods. It has been adopted in studies such as Regina et al. (2020).
- **BEC-Guard:** A proprietary dataset from Barracuda Networks comprising roughly 7,000 labeled BEC attack emails. This dataset was introduced by Cidon et al. (2019) as part of the BEC-Guard system for anomaly-based fraud detection.

- **Russian & English Emails:** A private multilingual corpus containing 2,308 genuine and simulated emails from 50 authors. It was used by Vorobeva et al. (2021) to assess BEC detection models across languages.

To synthesize the wide range of countermeasures proposed in the literature, Table 3.3 provides a consolidated summary of Business Email Compromise (BEC) studies, categorizing each work according to the types of solutions addressed. The table distinguishes between non-technical approaches—such as awareness training, human verification, and governance policies—and technical solutions, including machine learning, natural language processing, checksums, cryptographic techniques, intrusion detection systems, and firewalls. This classification enables a clearer comparison of the methodological diversity and focus areas within existing BEC research.

3.3 Authorship Verification (AV) Literature

AV asks whether two texts were produced by the *same* writer when the set of possible authors is open. The task underpins a wide range of high-stakes applications—from forensic linguistics and plagiarism detection to continuous user authentication in cyber-defence systems. In Business-E-mail-Compromise (BEC) scenarios, AV is especially valuable: attackers obfuscate malicious intent by borrowing the lexical habits and tonal cues of executives or suppliers, defeating rule-based spam filters that look only for links, attachments, or header anomalies.

Historically, progress in AV has mirrored the broader evolution of NLP. Tabel 3.4 (page 33) compiles representative studies across three eras—*traditional*, *hybrid*, and *modern*—and highlights the steady move from handcrafted stylometry towards deep, context-rich representations.

3.3.1 Traditional Era: Hand-engineered Stylometry

Early systems treated style as a stable set of surface cues. Common feature spaces included character or word n-grams, function-word frequencies, punctuation profiles, and vocabulary-richness indices Ruder et al. (2016); Abbasi and Chen (2005). Simple distance measures such as the *out-of-place* n-gram metric Kešelj et al. (2003) or Burrows's Delta Burrows (2002) were paired with linear classifiers—most notably SVMs and Naïve Bayes—to yield respectable accuracy on homogeneous corpora. However, these models degraded sharply when topic, genre, or document length varied, a weakness that limits their usefulness for short, domain-specific e-mail.

3.3.2 Hybrid Era: Statistical Learning with Shallow Embeddings

To bridge the gap between rigid stylometry and fully learned representations, researchers began to combine lightweight feature extraction with statistical learning. Profile-based dissimilarity

TABLE 3.3: Summary of BEC studies and methods.

Study	Non-Technical Solutions			Technical Solutions					
	Awareness Training	Human Verification	Policiesand Guidelines	Machine Learning	NLP Methods	Checksum	Cryptography	Intrusion Detection	Firewall
Mansfield-Devine (2016)	✓								
Zweighaft (2017)	✓								
FBI (2021)		✓							
Ross (2018)	✓								
Benaroch (2018)			✓						
Nehme and George (2018)	✓								
Meyers (2018)			✓						
Cidon et al. (2019)				✓					
Siadati (2019)								✓	
Maleki (2019)				✓					
Binks (2019)	✓								
Acar et al. (2019)			✓						
Baby et al. (2019)									✓
Kurematsu et al. (2019)				✓					
Awah Buo (2020)			✓						
Haddon (2020)									✓
Teerakanok et al. (2020)						✓			
Siadati et al. (2020)			✓						
Shahrivari et al. (2020)			✓						
Aparna et al. (2021)									✓
Vorobeva et al. (2021)				✓	✓				
Wickline (2021)				✓					
Susanti et al. (2023)			✓						
Ogwo-Ude (2023)			✓						
Almutairi et al. (2023)				✓	✓				
Papathanasiou et al. (2023)	✓								
Brabec et al. (2023)				✓					
Papathanasiou et al. (2024)							✓		
Lazarus (2024)	✓								

measures Potha and Stamatatos (2014) and compression-distance kernels Halvani et al. (2017) removed manual feature weighting, boosting robustness across languages while remaining computationally light. Recurrent architectures such as multi-headed RNN auto-encoders Bagnall (2015) captured sequential context, but still required elaborate hyper-tuning and struggled with very short texts typical of BEC mail.

3.3.3 Modern Era: Deep and Transformer-based Models

State-of-the-art systems now view AV as a representation-learning problem. Siamese and contrastive networks learn to project text pairs into a latent space where "same-author" instances cluster tightly while "different-author" pairs repel Araujo-Pino et al. (2020); Tyo et al. (2021). Transformer encoders supply the linguistic backbone: BERTAA fine-tunes BERT to push cross-topic AUC to 0.89 on Enron/IMDb data Fabien et al. (2020); Longformer adds global-window attention to handle 4 000-token novels with a 5-point accuracy boost over vanilla BERT Ordoñez (2020). Open-set variants with XLNet gating halve false-positive rates when previously unseen authors appear Peng (2021). The cost of these gains is increased model size, inference latency, and a risk of topic leakage.

3.3.4 Comparison of AV Methods

To better understand the progress in AV research, Table 3.4 summarizes the key methods, their advantages, and limitations.

TABLE 3.4: Representative studies on authorship-verification (AV) techniques.

Study	Era	Doc. type†	Model / Technique	Headline result	Main strengths	Key limitations
Kešelj et al. (2003)	Traditional	E-mail, *short*	Out-of-place character n-gram distance	>90% accuracy on Enron	Fast, no training	Breaks with topic drift
Koppel and Schler (2004)	Traditional	Essays, *long*	Function-word SVM ("unmasking")	≈85% on essays	Interpretable weights	Poor cross-domain generalization
Potha and Stamatatos (2014)	Traditional	Mixed, *short*	Profile-based distance (PAN-AV'14 winner)	Top ranked, good precision	Lightweight	Low recall on long docs
Halvani et al. (2017)	Hybrid	Mixed	Compression-distance kernel	+8 F1 over Delta on PAN-AV'17	No features; language-agnostic	High memory use
Bagnall (2015)	Hybrid	Blogs, *long*	Multi-headed RNN autoencoder	≈88% correct	Sequential context learned	Expensive training
Araujo-Pino et al. (2020)	Modern	PAN email, *short*	Siamese CNN on char n-grams	≈80% in PAN-20	Learns similarity	Sensitive to padding
Fabien et al. (2020)	Modern	Enron / IMDb	BERTAA (BERT + MLP)	AUC ≈0.89	Deep context; minimal features	Large, slow; may overfit topics
Ordoñez (2020)	Modern	Novels, *long*	Longformer	+5% over BERT on 4k tokens	Handles long input	Depends on partitioning
Tyo et al. (2021)	Modern	Corp. Email, *short*	Siamese RoBERTa	78% on corporate email	Strong for formal text	Needs careful tuning
Peng (2021)	Modern	Blogs, mixed	XLNet + open-set gating	Halves false positives on new authors	Handles unseen authors	Very compute-intensive

†Short ≈ 500 tokens or fewer; Long = multi-paragraph/multi-doc.

3.4 Multi-Task Learning (MTL) Literature

Multi-Task Learning (MTL) has proven effective in various NLP applications where related tasks can reinforce one another. For example, Plaza-Del-Arco et al. (2021) demonstrated that sharing a BERT encoder across hate-speech detection, sentiment analysis, and emotion classification improved performance on low-resource hate-speech benchmarks by leveraging affective cues. Similarly, Qu et al. (2022) combined text–hashtag semantic matching with informativeness detection to identify hashtag hijacks in social media, reducing false positives by forcing the shared encoder to learn both topical alignment and pragmatic intent.

In the domain of deception, Kumari et al. (2021) used MTL to fuse fake-news detection with novelty detection and emotion recognition, showing that auxiliary "novelty" and "emotion" tasks improved overall accuracy. Likewise, Choudhry et al. (2022) jointly predicted emotion and rumor legitimacy, achieving better cross-domain generalization. Jing et al. (2021) extended MTL to multimodal fake-news classification by integrating text, images, and comment-sentiment variance—though its reliance on social-media metadata limits direct applicability to email.

3.5 Literature Critique and Research Gaps

A critical review of the literature on BEC, AV, and MTL reveals consistent gaps that shape the scope and direction of this thesis. While important advances have been made, the state of the art remains fragmented and limited in several respects.

From a BEC perspective, most technical approaches rely heavily on metadata analysis (e.g., headers, sender reputation, SPF/DKIM checks). These signals, while useful in detecting spoofed domains, are ineffective in account-compromise scenarios where malicious emails are sent from legitimate infrastructure. Content-driven approaches are comparatively underexplored, and those that exist often emphasize shallow semantic features without deeper stylistic analysis. On the non-technical side, measures such as user training and policy frameworks are frequently proposed but lack rigorous empirical evaluation and are rarely integrated with technical detection systems.

AV has been widely studied in domains such as social media and academic texts, with both stylometric and neural approaches demonstrating promising results. However, AV has seldom been applied in enterprise email contexts, despite its direct relevance to impersonation-based attacks. Existing work typically focuses on closed-set author identification or small-scale corpora, overlooking adversarial conditions such as mimicry, where attackers intentionally imitate writing styles.

MTL has proven effective across many NLP tasks by enabling shared representations, improving generalisation, and reducing inference costs. Yet, its use in cybersecurity is still limited, and no prior research has explored combining BEC detection with AV in a single framework. This leaves

unexplored the potential synergies between semantic fraud detection and stylistic author profiling, which could strengthen defences against impersonation-driven attacks.

Based on the analysis, four core gaps in the existing literature are identified:

- **Lack of task-specific taxonomies:** Unlike phishing or spam, BEC lacks a standardised classification scheme. This hinders consistent comparison across studies and obscures the true coverage of existing defence strategies.
- **Over-reliance on metadata artifacts:** Most BEC detection methods depend on mutable features such as email headers or domain verification. These approaches fail under account-compromise conditions where attackers send messages from authentic infrastructure.
- **Absence of enterprise-focused AV research:** Existing AV studies rarely address enterprise email communication or adversarial settings. Few works consider mimicry attacks, where attackers emulate stylistic patterns of legitimate users, leaving a critical gap for BEC defence.
- **Limited exploration of joint-task learning:** While MTL is established in NLP, it has not been applied to cybersecurity tasks such as BEC and AV. No prior work has examined how semantic and stylistic tasks can be modelled together in a unified framework to improve accuracy and efficiency.
- **Dataset scarcity for impersonation-based BEC:** Public datasets such as Enron and TREC support general email classification but do not capture adversarial impersonation. There is a lack of realistic, labelled corpora that include both legitimate and mimicry-style BEC attacks.

Bridging the Gaps

This thesis directly addresses these gaps:

- *Chapter 5* introduces a five-axis taxonomy covering anatomy, methodology, target, counter-measure, and challenge, providing a structured lens for analysing BEC.
- *Chapter 6* details the creation of a synthetic dataset tailored to impersonation-based BEC, incorporating semantic deception and stylistic mimicry for realistic evaluation.
- *Chapter 7* presents a transformer–BiLSTM detector that focuses on lexical, syntactic, and semantic cues, reducing dependence on mutable metadata.
- *Chapter 8* develops BiBERT-AV, a Siamese network designed for enterprise emails, which learns stylistic signatures and detects inconsistencies under mimicry scenarios.
- *Chapter 9* proposes a novel MTL framework that jointly models BEC detection and AV, leveraging shared representations to enhance performance and reduce inference cost.

3.6 Chapter Summary

This chapter surveyed the current state of Business E-mail Compromise (BEC) defence in order to answer *SRQ1: What approaches currently exist for detecting BEC attacks, and what are their respective strengths and limitations?* The literature divides naturally into technical and non-technical counter-measures. Technical proposals range from rule-based checksums and header verifications to recent transformer-based classifiers that inspect message content; non-technical measures encompass employee-awareness training, human verification steps, and governance policies for high-value transactions. A comparative table showed that, while machine-learning methods now dominate academic work, many commercial products still depend almost exclusively on metadata signals (SPF, DKIM, IP reputation).

In parallel, the review catalogued the publicly and privately available datasets (Enron, TREC, BEC-Guard, simulated corpora) and highlighted persistent data issues: scarcity of labelled BEC examples, class imbalance, and privacy constraints.

The absence of a shared classification scheme is the most fundamental barrier, because it prevents cumulative progress and obscures the true coverage of existing defences. The next chapter therefore introduces a five-axis BEC taxonomy that standardises terminology across anatomy, methodology, target, counter-measure, and detection challenge. This taxonomy provides the conceptual scaffold on which the thesis builds its subsequent detection models and evaluation protocols.

Chapter 4

Research Methodology

This chapter presents the research methodology employed to address the thesis objectives and sub-research questions. The thesis adopts a mixed-methods approach, integrating quantitative experimentation with qualitative thematic analysis. As discussed by Creswell and Clark (2017), this form of methodological triangulation enhances the validity, reliability, and interpretive richness of the findings. Figure 4.1 outlines the four research phases.

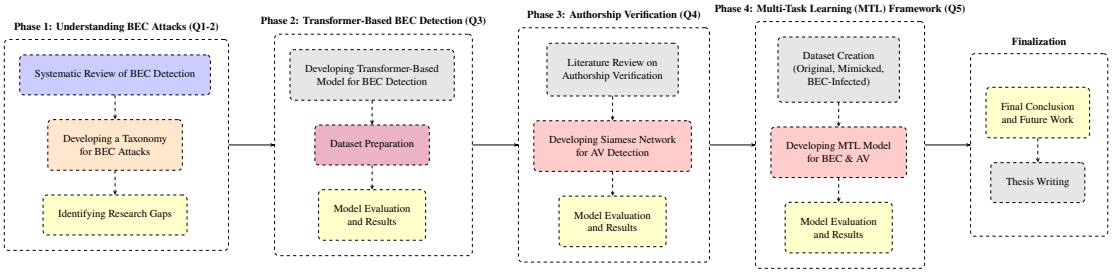


FIGURE 4.1: Research Phases

4.1 Research Methods Overview

Research methods were selected based on the nature of each sub-research question (SRQ) and the characteristics of the data involved. This thesis primarily relies on:

- **Quantitative methods:** for model development, experimentation, and evaluation.
- **Qualitative methods:** for literature review, taxonomy construction, and gap identification.
- **Mixed-methods integration:** for triangulation and methodological complementarity.

4.1.1 Quantitative Methods

Quantitative methods were central to Phases 2–4, involving experimental validation of transformer-based models across multiple tasks. These experiments adhered to the following process:

1. Defining research questions and experimental goals.
2. Selecting datasets, baseline models, and performance metrics.
3. Implementing models (BEC detection, AV, and MTL).
4. Conducting comparative evaluation.
5. Interpreting and documenting results.

4.1.2 Qualitative Methods

Qualitative analysis was employed during Phase 1 to synthesize existing literature and extract conceptual insights. Thematic analysis informed the development of a five-axis taxonomy for BEC attacks and highlighted underexplored areas Taylor (2005).

4.1.3 Mixed-Methods and Triangulation

An embedded sequential mixed-methods design was followed Lister (2005). Qualitative insights from Phase 1 informed subsequent quantitative experimentation. Triangulation across methods and datasets enhanced the robustness and credibility of findings Cohen (2007); Runeson and Höst (2009).

4.2 Research Paradigm

This thesis adopts a pragmatic research paradigm, allowing for methodological flexibility and prioritizing practical problem-solving in the cybersecurity context Creswell and Clark (2017). This aligns with the thesis goal of developing deployable and interpretable models for BEC detection and authorship verification.

4.3 Detailed Research Phases

The thesis is organized into four sequential phases, each mapped to a sub-research question.

4.3.1 Phase 1: Systematic Understanding of BEC

SRQ1: What approaches currently exist for detecting BEC attacks, and what are their respective strengths and limitations?

SRQ2: How can BEC attacks be systematically categorized to support effective detection and prevention strategies?

A systematic literature review was conducted following PRISMA guidelines Kitchenham and Charters (2007). A novel taxonomy was developed and validated through real-world case studies, framing the research scope and identifying design constraints for the models proposed in later phases.

4.3.2 Phase 2: Transformer-Based BEC Detection

SRQ3: How effective are transformer-based classifiers for phishing text-based attacks, and to what extent do they generalise to impersonation-driven phishing text-based attacks when only email body content is available?

This phase involved designing a transformer–BiLSTM hybrid architecture, comparing it to traditional baselines on real and synthetic datasets. Emphasis was placed on detecting impersonation attacks without reliance on metadata.

4.3.3 Phase 3: Siamese Network for Authorship Verification

SRQ4: How do transformer-based Siamese networks perform in authorship verification of business emails compared to traditional stylometric methods?

A BiBERT-AV architecture was proposed, using paired input structures to assess stylistic similarity. Evaluation was conducted on mimicry and non-mimicry scenarios to validate the model's robustness.

4.3.4 Phase 4: Multi-Task Learning (MTL) Framework

SRQ5: How does integrating BEC detection and authorship verification in a single system affect overall accuracy and operational cost?

This phase introduced a unified MTL framework with a shared encoder and task-specific heads. Joint training was expected to improve generalization and reduce computational redundancy. The model was benchmarked against sequential transfer learning and single-task baselines.

4.4 Research Question–Method–Outcome Mapping

Table 4.1 summarizes the alignment between each research phase, its associated research question, methodological approach, and key outcome.

TABLE 4.1: Mapping of research questions, methods, and outcomes

Phase	**RQ**	**Methods**	**Outcomes**
1	SRQ1–2	Systematic review, thematic analysis	BEC taxonomy, research gaps
2	SRQ3	Transformer-based experimentation	BEC detection model, comparative results
3	SRQ4	Siamese transformer network	AV model, mimicry robustness
4	SRQ5	MTL training, ablation studies	Joint framework, performance improvement

Dataset Note. A custom synthetic dataset was developed to support Phases 2–4. Details are provided in **Chapter 6**.

4.5 Unified Experimental Setup

To ensure comparability across experiments, a standardized setup was adopted as follow:

4.5.1 Implementation

All models were implemented in Python (v3.9–3.11) using PyTorch Paszke et al. (2019), Hugging Face Transformers Wolf et al. (2020), and scikit-learn Pedregosa et al. (2011). BERT variants were used with default tokenizers and maximum input length of 256.

4.5.2 Hardware

Most experiments were run on an NVIDIA A100; initial AV runs used an NVIDIA P100.

4.5.3 Data Splitting and Preprocessing

All datasets were split into 70% training, 10% validation, and 20% testing. This split ratio is widely used in supervised learning research to ensure a sufficient volume of training samples for deep models, while maintaining reliable validation and unbiased test sets Deng and Liu (2018).

The 10% validation portion is used for early stopping and hyperparameter tuning, and the final evaluation is conducted on the held-out 20% test set.

Email body preprocessing included:

- Removal of headers, signatures, URLs, HTML, and non-alphabetic characters.
- Lowercasing and tokenization using BERT's tokenizer.
- Dynamic padding and truncation.

4.5.4 Training Settings

All models were trained for up to 10 epochs using:

- AdamW optimizer with learning rate 2×10^{-5}
- Batch size: 32
- Dropout: 0.3
- Early stopping on validation loss

4.5.5 Evaluation Metrics

Standard classification metrics were applied throughout:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}, \qquad \text{Precision} = \frac{TP}{TP + FP},$$

$$\text{Recall} = \frac{TP}{TP + FN}, \qquad \text{F1-score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}},$$

$$\text{False Positive Rate (FPR)} = \frac{FP}{FP + TN}, \qquad \text{AUC} = \int_0^1 \text{TPR}\left(\text{FPR}^{-1}(x)\right)\, dx.$$

For authorship verification, predictions were based on a decision threshold applied to similarity scores. Metrics were computed accordingly.

4.6 Chapter Summary

This chapter described the research paradigm, methodological phases, and experimental design. A mixed-methods approach underpins the thesis, combining systematic review, transformer-based experiments, and integrated evaluation metrics. The next chapter presents the transformer-based BEC detection experiments.

Chapter 5

The Proposed BEC Taxonomy

Introduction

BEC differs markedly from traditional phishing in both strategy and execution. Rather than relying on generic deception or malicious links, BEC campaigns employ tailored impersonation, social-engineering scripts, and linguistically crafted requests that appear credible to their targets. These characteristics vary widely across incidents, complicating efforts to compare studies, share threat intelligence, or develop robust, generalisable defences.

As discussed in Chapter 3, previous literature has explored isolated aspects of BEC—such as impersonation tactics or financial fraud vectors—but lacks a unified, structured taxonomy that systematically captures the full range of observed behaviours. The absence of a standardised classification framework limits both academic progress and practical application in the field.

The aim of this chapter is to fill that gap by proposing a comprehensive taxonomy specifically designed for BEC. This taxonomy supports systematic categorisation, enhances conceptual clarity, and provides a foundation for designing better detection models and response strategies.

Objectives of this chapter:

- Introduce a five-axis taxonomy that captures the full complexity of BEC incidents, including their forms, tactics, intended targets, mitigation strategies, and detection challenges.
- Facilitate a deeper understanding of BEC behaviour to support the design of more effective prevention and detection mechanisms.
- Offer a structured reference for future research and operational defence systems in the domain of e-mail-based fraud.

[0]This chapter is based on the publication: Almutairi, A. M., Kang, B., & Al Hashimy, N. (2024). *Business Email Compromise: A Comprehensive Taxonomy for Detection and Prevention*. In *Proceedings of the 7th International Conference on Information Science and Systems (ICISS '24)*, pp. 49–54.

This contribution directly addresses **SRQ2:** *How can Business Email Compromise (BEC) attacks be systematically categorized to support effective detection and prevention strategies?*

To answer this question, the chapter introduces a five-axis taxonomy derived from patterns in the systematic literature review (Chapter 3). The axes are deliberately orthogonal: *who* is being impersonated (persona) is distinct from *what* is being sought (objective), *how* it is attempted (operational technique), *who* is pressured to act (target), and *how* it is mitigated (countermeasures).

1. **Persona (Pretext):** the claimed sender identity used to confer authority or familiarity (e.g., internal VIP, manager/colleague, vendor/partner, authority/regulator).
2. **Objectives (Outcomes):** the business end-goal requested from the recipient (e.g., payment diversion, payroll/benefits diversion, data theft, commodity fraud, process abuse).
3. **Operational Techniques:** the concrete tactics used to execute the scheme (e.g., identity deception without ATO, account takeover/EAC, conversation manipulation, payment-instruction alteration, alternative-channel handoff).
4. **Targets:** the recipient/approver roles expected to act (e.g., AP/Finance/Treasury, executives and assistants, HR/Payroll, vendors/partners, IT/helpdesk).
5. **Countermeasures:** technical and process controls aligned to the above axes (e.g., SPF/DKIM/DMARC with alignment, MFA and conditional access, OAuth restrictions, EAC detection, content analytics including authorship verification, URL/attachment protection, out-of-band verification, dual approval/segregation of duties, vendor-bank verification, role-tailored awareness and escalation).

The following sections elaborate each axis in turn.

5.1 Classification by Persona (Pretext)

The *claimed sender identity* used to establish authority or familiarity:

- **Internal VIP** (e.g., CEO/CFO), **manager/colleague**.
- **Vendor/customer/partner** in the supply chain.
- **Authority/regulator** (e.g., legal, auditor).

5.2 Classification by Objectives (Outcomes)

The *business end-goal* requested from the recipient:

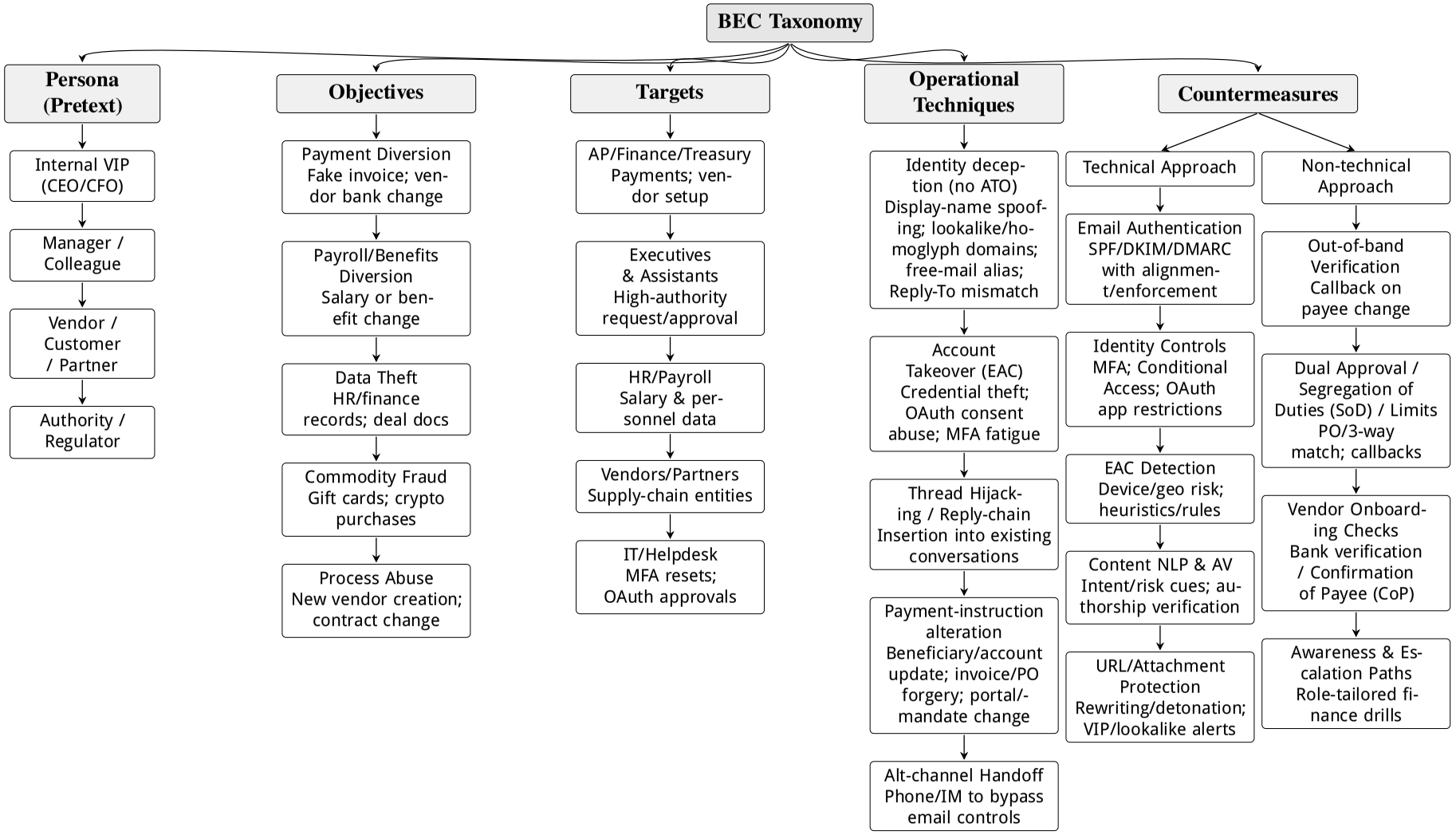


FIGURE 5.1: BEC taxonomy.

- **Payment diversion** (e.g., fake invoice; vendor bank change).
- **Payroll/benefits diversion** (salary redirection; benefits changes).
- **Data theft** (HR/finance records; deal documents).
- **Commodity fraud** (gift cards; crypto purchases).
- **Process abuse** (new-vendor creation; contract/approval changes).

5.3 Classification by Operational Techniques

Concrete tactics used to prosecute the scheme (avoiding generic labels that apply to most BEC):

- **Identity deception (no ATO):** display-name spoofing; lookalike/homoglyph domains; free-mail aliases; `Reply-To` mismatch.
- **Account takeover (EAC):** credential theft; OAuth consent abuse; MFA fatigue/prompt bombing.
- **Conversation manipulation:** thread hijacking/reply-chain insertion; urgency/confidentiality framing; staged approvals.
- **Payment-instruction alteration:** beneficiary/account update; invoice/PO forgery; portal/mandate change.
- **Alternative-channel handoff:** phone/IM continuation to bypass email controls.

5.4 Classification by Targets

Recipient/approver roles expected to act:

- **AP/Finance/Treasury** and **Procurement/vendor master**.
- **Executives** and **executive assistants**.
- **HR/Payroll**.
- **Vendors/partners** (supply chain).
- **IT/Helpdesk** (e.g., MFA resets; OAuth approvals).

5.5 Classification by Countermeasures

A layered posture combines technical and organisational controls aligned to the above axes.

Technical

- **Email authentication:** SPF/DKIM/DMARC with alignment/enforcement (optionally MTA-STS/BIMI).
- **Identity controls:** MFA; conditional access; OAuth application restrictions.
- **EAC detection:** device/geo risk, heuristics, mailbox-rule/behavioural signals.
- **Content analytics:** NLP-based intent/risk cues and authorship verification; URL/attachment protection.

Organisational

- **Out-of-band verification** and **dual approval/segregation of duties**.
- **Vendor onboarding/changes:** bank verification; Confirmation of Payee.
- **Role-tailored awareness** and clear escalation/reporting paths.

5.6 Validation of the Taxonomy

Descriptive power is a widely used approach for validating taxonomies Nickerson et al. (2013) and has been applied to phishing Garera et al. (2007) and threat intelligence Tounsi and Rais (2018) in cybersecurity contexts. We therefore assess whether the proposed axes cleanly and completely describe real BEC incidents.

5.6.1 Case Study 1: Treasure Island Homeless Charity

Overview. In June 2021, Treasure Island, a San Francisco–based homelessness charity, suffered a BEC loss of $625,000. Attackers gained access to the bookkeeper's mailbox and manipulated a legitimate vendor invoice, leading to funds being diverted to an attacker-controlled account Tessian (2021).

Taxonomy mapping.

- **Persona (Pretext):** Vendor/partner (invoice origin); internal colleague (bookkeeper) context enabled by EAC.
- **Objectives (Outcomes):** Payment diversion (beneficiary/bank-account change on a legitimate invoice).
- **Operational Techniques:** Account takeover (EAC); conversation manipulation via thread hijacking/reply-chain insertion; payment-instruction alteration (beneficiary/account update).

- **Targets:** AP/Finance (bookkeeping/treasury staff executing payments).
- **Countermeasures:** Identity controls (MFA, conditional access) and EAC detection; content analytics for payment-intent cues; out-of-band payee verification and dual approval/segregation of duties; vendor-bank verification (e.g., Confirmation of Payee).

5.6.2 Case Study 2: Insurance Broker Firm

This case study applies the proposed taxonomy to a real incident involving an insurance broker, as documented by Kroll (2021). The broker's environment was compromised and then used to solicit a fraudulent payment from a client.

Overview. An attacker obtained broker credentials via phishing and, six weeks later, used the compromised mailbox to request that a client redirect nearly £300,000 to an alternative account. The attempt was detected before funds were transferred Kroll (2021).

Taxonomy mapping.

- **Persona (Pretext):** Vendor/partner (the broker, communicating with its client).
- **Objectives (Outcomes):** Payment diversion (alternate beneficiary/bank account).
- **Operational Techniques:** Account takeover (EAC) enabled by credential harvesting; conversation manipulation (reply-chain use within an existing relationship); payment-instruction alteration (beneficiary/account update).
- **Targets:** External partner—client AP/Finance team responsible for payment execution.
- **Countermeasures:** Identity controls (MFA, conditional access, OAuth restrictions) and EAC detection; content analytics for payment-intent cues; process controls including out-of-band payee verification, dual approval/segregation of duties, and bank-account verification (e.g., Confirmation of Payee); post-incident forensics for scoping and hardening.

5.7 Chapter Summary

This chapter introduced an orthogonal, five-axis taxonomy for Business Email Compromise comprising *Persona (Pretext)*, *Objectives (Outcomes)*, *Operational Techniques*, *Targets*, and *Countermeasures*. This taxonomy provides the conceptual foundation for the detection models developed in subsequent chapters.

Chapter 6

Synthetic Dataset Creation

6.1 Introduction

Business Email Compromise (BEC) research faces an immediate obstacle: there is no open collection of genuine BEC e-mails. Incident reports and raw messages are typically protected by non-disclosure agreements or privacy law, preventing their release and, by extension, hindering reproducible experimentation. Multiple attempts to obtain real-world samples—from researchers, security vendors, and enterprise contacts—were unsuccessful due to the legal sensitivity and confidentiality of such incidents.

Because no authentic BEC datasets are openly available, researchers typically rely on general-purpose email corpora such as Enron or anti-spam benchmarks (CEAS, TREC, LingSpam). Although these collections are sizeable, they weren't built to capture the impersonation tactics, organisational role dynamics, and high-stakes payment pressure that define BEC attacks. Table 6.1 lists studies that have used this workaround.

TABLE 6.1: Examples of studies that substituted public email sets for real BEC data.

Study	Corpora used	Focus
Maleki (2019)	Enron folders	Stylistic BEC detection
Cidon et al. (2019)	Live mail + Enron	Production BEC filter
Xiao and Jiang (2020)	TREC'06p, fraud letters	Spam/phish filtering
Brabec et al. (2023)	TREC'07 + custom phishing	Modular BEC detection
Alguliyev et al. (2024)	LingSpam, Enron-Spam, TREC'07	BERT/BiGRU BEC study

To address this gap, we built a *synthetic* email dataset that mirrors common BEC attack patterns. The generation process started with a small set of real BEC incidents and then fine-tuned a LLM model on the Enron corpus and real BEC samples so the output sounds like ordinary business mail. This gives us the main wording, structure, and impersonation hints seen in real attacks, even though the full operational context is not present.

Our aim is to support research in both BEC detection and authorship verification and includes three major components:

- The three most prevalent BEC attack types, as identified in our literature review and taxonomy chapter: *Bogus Invoice*, *CEO Fraud*, and *Account Compromise*.
- Benign emails that mimic the writing style of real users, to simulate challenging authorship verification cases.
- Impersonation-based BEC messages, which combine deceptive content with style mimicry—arguably the most difficult class of BEC threats to detect.

The remainder of this chapter addresses the dataset construction process in response to the data availability challenges outlined above. Section 6.2 outlines the methodology used to build a task-aligned email corpus. Section 6.3 discusses the ethical and legal considerations guiding its development. Finally, Section 6.4 reflects on the dataset's limitations and proposes directions for future extension.

6.2 Methodological Pipeline

To address the scarcity of accessible BEC datasets, we developed a structured pipeline for generating a synthetic corpus that combines semantic deception and stylistic impersonation. This pipeline integrates real-world seeds, transformer-based text generation, and multiple layers of quality control. Figure 6.1 illustrates the structured pipeline that underpins this dataset in nine-stage process:

6.2.1 Step 1: Data Sources

We collected 21 real BEC emails from publicly available sources, including threat intelligence reports, academic papers, and security blogs. These samples served as seed examples for generating synthetic BEC messages.

For benign communication and authorship modelling, we used the Enron Email Dataset, a widely adopted corporate email corpus. We selected five authors who had each sent more than 1,000 emails to ensure sufficient data and stylistic consistency. Emails were extracted from the respective sender folders.

6.2.2 Step 2: Preprocessing

All emails were preprocessed using standard text-cleaning steps. This included removing headers, signatures, URLs, and HTML tags. The text was then lowercased, and punctuation was normalized to prepare the content for model input and subsequent stylistic analysis.

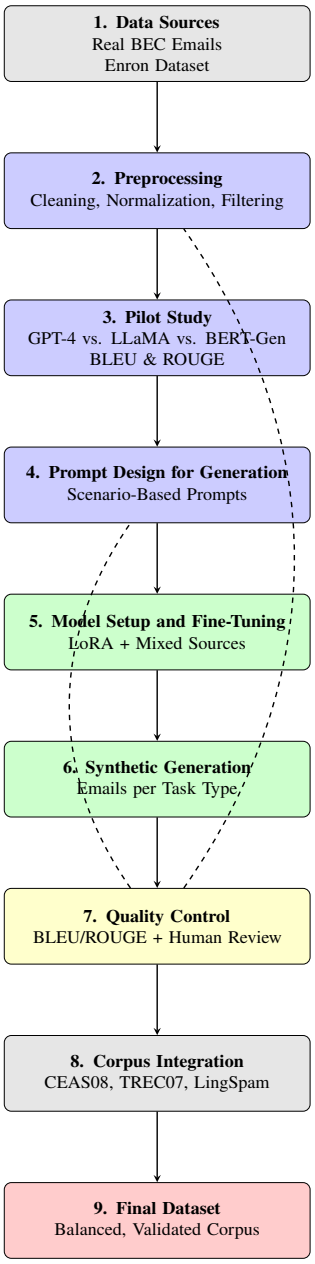


FIGURE 6.1: Overview of the dataset generation pipeline.

6.2.3 Step 3: Pilot Study – Model Selection

To identify the most suitable language model for generating realistic BEC-style emails, we conducted a pilot study comparing GPT-4, LLAMA, and BERT-generation models. The comparison was based on 21 seed emails. Our goal was to evaluate which model best replicates the lexical and structural patterns observed in real messages.

We used BLEU and ROUGE-L scores to measure the similarity between the generated emails and their reference seed emails:

- **BLEU (Bilingual Evaluation Understudy)** — Measures n-gram precision, indicating how well the generated text preserves the original lexical content.
- **ROUGE-L (Recall-Oriented Understudy for Gisting Evaluation)** — Measures the longest common subsequence between texts, reflecting structural similarity.

For each seed email, we generated 50 variants using each model and calculated the average BLEU and ROUGE-L scores. This provided a consistent way to compare how closely each model could emulate the style and structure of the reference emails.

Table 6.2 shows the results across increasing numbers of seed prompts. LLaMA consistently achieved the highest scores, demonstrating better lexical fidelity and structural similarity than the other models.

TABLE 6.2: Average BLEU and ROUGE-L scores across seed ranges (higher is better).

# Seeds	Model	BLEU	ROUGE-L
0–5	GPT-4	0.0004	0.030
	LLaMA	**0.046**	**0.038**
	BERT-Gen	0.006	0.020
0–10	GPT-4	0.0015	0.034
	LLaMA	**0.046**	**0.043**
	BERT-Gen	0.008	0.024
0–15	GPT-4	0.0033	0.037
	LLaMA	**0.026**	**0.048**
	BERT-Gen	0.007	0.028
0–21	GPT-4	0.0041	0.039
	LLaMA	**0.011**	**0.053**
	BERT-Gen	0.007	0.032

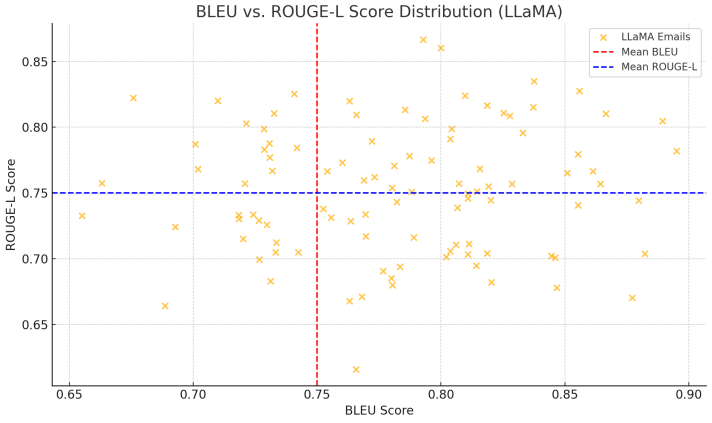


FIGURE 6.2: Distribution of BLEU and ROUGE-L scores for LLaMA.

The decision to use LLaMA was based on its consistent performance across all seeds using well-established metrics. These scores provided a practical and scalable way to estimate how well each model could replicate the lexical and structural characteristics of real BEC messages.

6.2.4 Step 4: Model Setup and Fine-Tuning

We fine-tuned LLaMA using Low-Rank Adaptation (LoRA) for parameter-efficient adaptation. Fine-tuning was carried out separately for two subtasks: (a) BEC scenario generation (from the 21 seed frames), and (b) authorship mimicry (per-author, Enron-based) (from the 1000 seed frames).

Parameter-efficient setup. Only adapter weights were trained; base model weights were frozen. LoRA targeted attention projections (`q_proj`, `k_proj`, `v_proj`, `o_proj`).

Table 6.3: Fine-tuning configuration.

Setting	Value
Max input length	256 tokens
Batch size	16
Optimizer & LR	AdamW, 2×10^{-4} (adapters only)
Weight decay / betas	0.0 / (0.9, 0.999)
Epochs / early stop	3 (BEC), 2 (per-author mimicry) / dev perplexity
LoRA rank r / α / dropout	8 / 16 / 0.1

6.2.5 Step 5: Prompt Design for Generation

After fine-tuning LLaMA, we designed a set of hand-crafted prompts to reflect common Business Email Compromise (BEC) scenarios. These prompts were manually written based on insights from real BEC incidents and crafted to simulate authentic business communication.

All synthetic emails in the BEC and authorship mimicry subsets were generated using these manually curated prompts. Each prompt was written to simulate either a benign corporate email or a malicious message aligned with known BEC strategies. For example:

> "`Compose a formal email requesting a wire transfer due to a last-minute invoice adjustment.`"

For style-controlled samples, prompt templates included explicit instructions related to both the communicative goal (e.g., request for payment) and the stylistic identity of the sender (e.g., "in the style of Author 3"). This allowed the generated emails to exhibit both semantic relevance and stylistic fidelity.

We adopted a traditional manual prompt engineering approach—commonly referred to as hand-crafted prompting—which prioritizes interpretability and control over scalability. This contrasts with prompt optimization techniques that automatically refine prompt content based on objective functions or feedback signals.

6.2.6 Step 6: Synthetic Email Generation

Following fine-tuning, the LLaMA model was used to generate large-scale synthetic samples across multiple scenarios, including BEC attacks, authorship mimicry, and impersonation. For each scenario, prompts were designed to produce diverse outputs by varying tone, structure, and wording, while maintaining the core intent of the message.

The resulting dataset consists of four main subsets:

- **BEC Dataset Generation**, generated from real-case seeds and designed to reflect common threat scenarios;
- **Integration of Phishing Corpora**, integrated from established corpora to introduce additional variation;
- **Authorship Mimicry Dataset**, where the model emulates the writing style of selected Enron authors;
- **Impersonation-Based BEC Generation**, which combine deceptive intent with author-specific style to simulate complex attack cases.

All subsets were generated using the fine-tuned LLaMA model, and describe as follow:.

6.2.6.1 (i) BEC Dataset Generation

We used 21 real BEC incidents—sourced from public research papers, blogs, and reports—as the foundation for generation. For each incident, the model was prompted to generate 50 realistic variants, resulting in a total of 1,050 synthetic BEC emails.

Prompts were crafted to reflect typical BEC themes, including:

- urgent financial requests sent by executive impersonators;
- follow-ups regarding vendor payment;
- confirmations of fictitious transactions or account changes.

Each prompt was written to preserve professional tone and embed common social engineering features (e.g., urgency, authority, and impersonation). Outputs with low BLEU or ROUGE scores were discarded and regenerated to maintain stylistic quality and coherence.

6.2.6.2 (iii) Authorship Mimicry Dataset

To simulate style-based threats, we constructed an authorship mimicry dataset by fine-tuning LLaMA individually on five Enron authors. Authors were selected based on having authored at least 1000 unique emails to ensure stylistic consistency.

Fine-tuning for each author was performed independently using LoRA. Prompt templates then guided the model to generate realistic business communications in each author's distinctive style, covering a range of corporate topics. This subset supports the authorship verification task by providing examples of stylistic imitation. A total of 5,000 emails (1,000 per author) were generated.

6.2.6.3 (iv) Impersonation-Based BEC Generation

This subset builds on the authorship mimicry task by generating emails that incorporate both an author's writing style and the deceptive intent of a BEC attack. To create this subset, we reused the author-specific LLaMA checkpoints trained in the mimicry stage. Prompts were designed to inject BEC-specific semantics into stylistically accurate messages. For example:

> "Write an email in the style of Author 2 requesting an urgent wire transfer to a vendor."

Prompts were written to preserve stylistic coherence while embedding typical BEC themes such as urgency, authority, and financial requests. Importantly, all content was fully generated by the model—no real emails were copied or reused—to maintain originality and avoid privacy concerns.

To promote diversity, prompt templates were rotated and varied in tone, phrasing, and scenario type (e.g., CEO fraud, bogus invoice, account compromise). Each of the five authors contributed 1,000 samples, resulting in 5,000 impersonation-based BEC emails in total.

Any weaknesses in novelty were mitigated by BLEU/ROUGE filtering in the quality control phase.

6.2.7 Step 7: Quality Control

The quality of the generated emails was assessed through a combination of automatic scoring and human validation. BLEU and ROUGE-L were used as initial filters to evaluate lexical and structural similarity to seed messages. Samples that fell below a predefined threshold were flagged and regenerated.

To further assess realism and plausibility, we conducted a manual evaluation on a subset of 50 randomly selected LLaMA-generated emails. This sample was drawn from the broader synthetic BEC set and was not stratified by specific attack types (e.g., CEO fraud or bogus invoice), ensuring a general assessment of generation quality.

Annotator Setup

Two PhD students with a cybersecurity background independently annotated a random sample of 50 LLaMA-generated emails using the checklist below. Before annotation, they reviewed two example items to align on definitions; no labels were shared during scoring.

Disagreement Resolution

Annotators scored independently. If their labels differed, a designated adjudicator (first PhD Student) reviewed the email and both checklists and applied the same four–item rule (≥ 3 items = BEC; ≤ 1 = non-BEC; exactly 2 = borderline). The adjudicator's decision was final, and the chosen label plus a one-line rationale were recorded for all 50 items.

Checklist Criteria

Each email body should be evaluated against the following four text-based indicators. An email meeting **three or more** of these criteria is likely a BEC attempt.

1. **Authority Language**
   Does the text claim to come from a high-ranking or authoritative role (e.g., "As the CFO, I need you to. . . " or "This is a directive from our CEO")?
2. **Unusual Financial Request**
   Does the message ask for an atypical or suspicious payment, wire transfer, invoice update, or change in payment instructions?
3. **Urgency or Secrecy Cue**
   Does the wording emphasize immediate action or strict confidentiality (e.g., "Act now, this cannot be shared," "This is urgent, do not forward")?
4. **Familiarity/Context Reference**
   Does the text invoke inside information, previous conversations, project names, or role-specific details that a real insider would know (e.g., "As discussed in last week's budget meeting. . . ")?

Usage: Each item was rated independently. An email was classified as a "convincing BEC" if it satisfied at least three of the four checklist items. Annotator disagreements were resolved using Disagreement Resolution; no third reviewer or arbitration process was used.

Agreement Metrics

Inter-rater agreement is reported in Table 6.4. We computed Cohen's κ to measure agreement beyond chance. The resulting value of approximately 0.69 indicates *substantial* agreement according to the Landis–Koch scale.

TABLE 6.4: Inter-rater agreement on "convincing BEC" classification ($N = 50$).

	Rater B		
Rater A	POSITIVE	NEGATIVE	Total
POSITIVE	29 (58%)	3 (6%)	32
NEGATIVE	4 (8%)	14 (28%)	18
Total	33 (66%)	17 (34%)	50

Relation to Automatic Filtering

All emails selected for human evaluation had already passed automatic filtering based on BLEU and ROUGE-L scores. The purpose of manual validation was to verify whether the automatically accepted samples exhibited realistic BEC features. No additional filtering or regeneration was performed based on human annotation; rather, this step served to confirm the plausibility and relevance of the retained outputs.

6.2.8 Step 8: Phishing Corpus Integration

To enhance the diversity of malicious email formats and support broader generalization, we integrated phishing messages from three well-known public corpora: CEAS, TREC, and LingSpam. These datasets were added to supplement the synthetic BEC messages with varied phishing styles and content.

Non-textual cues such as hyperlinks, attachments, and metadata were removed during preprocessing to ensure the dataset emphasizes linguistic and stylistic deception rather than surface-level indicators.

This step complements our broader objective: to cover a wide range of BEC attack scenarios—from basic impersonation with generic financial requests to advanced cases that involve mimicking the target author's writing style. By including both low-effort and highly personalized threats, the dataset supports robust model training across different levels of attacker sophistication.

6.2.9 Step 9: Final Dataset Packaging

The final dataset combines synthetic BEC samples, authorship mimicry emails, impersonation-based attacks, and real-world phishing and ham messages. This composition enables the training and evaluation of models that are sensitive to both textual content and writing style.

The dataset is designed to support two main tasks: Business Email Compromise (BEC) detection and AV. To capture realistic communication patterns, the dataset includes overlapping writing styles across benign and malicious emails. While such overlap improves realism, strict data partitioning was enforced to prevent label leakage and ensure reliable evaluation.

6.2.9.1 Subsets and Sources

Below is a summary of the sources, roles, and sample counts for each dataset subset:

- **Real BEC (21 samples)** — Manually collected from public incident reports, research articles, and cybersecurity blogs.
- **Synthetic BEC (1,050 emails)** — Generated from the real-case seeds using LLaMA, covering key BEC scenarios such as CEO Fraud, Bogus Invoice, and Account Compromise.
- **Phishing Corpora** — Messages from CEAS08, TREC07, LingSpam, and SpamAssassin were added to increase linguistic diversity and simulate basic phishing attacks.
- **Enron Authors (5,000 emails)** — Extracted from the sender folders of five prolific Enron authors and used both for fine-tuning and as benign examples in BEC detection.
- **Authorship Mimicry (5,000 emails)** — LLaMA-generated emails replicating the writing style of the five selected authors, without embedding malicious intent.
- **Impersonation-based BEC (5,000 emails)** — Generated by blending BEC attack scenarios with author-specific writing style, representing more sophisticated forms of deception.

This design allows for the exploration of both low-effort phishing detection and more complex impersonation-based threats. By combining semantic content and stylistic signals, the dataset provides a foundation for evaluating models under realistic adversarial conditions.

6.3 Ethical and Legal Considerations

No real individuals were explicitly modeled or referenced during the creation of this dataset. All generated emails were produced using synthetic identities and abstracted business scenarios. Prompts involving sensitive or legally ambiguous content—such as specific financial institutions,

employee names, or real transaction records—were deliberately excluded to avoid ethical or legal concerns.

The real BEC examples used as generation seeds were obtained from publicly available sources, including academic publications, security reports, and blogs. Only de-identified, paraphrased, or obfuscated content was used during prompt design to ensure that no personally identifiable information (PII) or confidential content was retained.

This synthetic dataset was created exclusively for academic research. It does not simulate, promote, or encourage malicious behavior, and its intended use is to support the development and evaluation of defensive technologies in cybersecurity.

Licensing requirements for all reused corpora—such as the Enron dataset, CEAS08, and other phishing corpora—were carefully reviewed and respected. Additionally, all LLaMA-based generation was performed using model versions that are explicitly permitted for research-only use.

6.4 Conclusion

This chapter has introduced a synthetic dataset tailored to the dual tasks of Business Email Compromise (BEC) detection and AV using NLP techniques. The dataset was constructed through a structured pipeline that combines real-case seed messages, prompt-based generation, LoRA fine-tuning on business-style corpora, and multi-stage quality control. By incorporating both semantic deception and stylistic mimicry, the dataset supports the evaluation of content-aware and style-sensitive models under a range of realistic threat conditions.

The main contributions of this dataset include:

- **Coverage of key BEC scenarios**—including CEO Fraud, Bogus Invoice, and Account Compromise—designed to reflect common attack patterns;
- **Style-consistent benign and impersonation messages**, enabling fine-grained evaluation of authorship-based defenses;
- **Integration of phishing corpora**, supporting generalization beyond narrowly defined BEC threats;

Limitations. Despite these contributions, several limitations should be acknowledged:

- **Limited scenario coverage:** The dataset focuses on three BEC archetypes; other forms such as payroll redirection, gift card scams, or supply chain fraud are not included and remain avenues for future extension.

- **Seed sample constraint:** The generation process was based on 21 real BEC messages. Although carefully curated, this small pool limits lexical and rhetorical diversity.
- **Synthetic fidelity:** While BLEU and ROUGE scores help filter outputs with poor surface overlap, they do not capture deeper semantic similarity or discourse coherence. More advanced metrics such as BERTScore or MAUVE will be considered in future releases.
- **Language and domain generalization:** The dataset is English-only and based on a specific corporate communication style (Enron). Multilingual and cross-industry generalization remains an open challenge.
- **Prompt diversity:** Although multiple prompts were used, they were handcrafted and not optimized via systematic methods (e.g., reinforcement learning or prompt tuning), potentially limiting variability in the generated content.

This dataset was developed to address the specific requirements of the research presented in this thesis. However, its modular structure and accompanying documentation offer a reusable foundation for broader investigations in content-based email security, authorship verification under adversarial conditions, and synthetic data generation methodologies. Future work may extend its applicability by introducing additional BEC archetypes, incorporating multilingual and cross-domain corpora, and leveraging more advanced generative models in collaboration with industry stakeholders. The final version of the dataset is publicly available for academic research purposes.[1]

[1] Synthetic BEC Dataset: `https://github.com/AmirahCoding/synthetic-bec-dataset`

Chapter 7

Transformer-Based Models for BEC Attack Detection

7.1 Introduction

Business Email Compromise (BEC) is among the most financially damaging forms of cyber-enabled fraud Internet Crime Complaint Center (IC3) (2023). Unlike classical phishing, BEC messages rarely include obvious indicators (malicious URLs, macros, or attachments) and often succeed through impersonation and organizational pretexting.

Chapters 3 showed that many defences rely on mutable metadata (SPF/DKIM alignment, IP reputation), and that content-only approaches are seldom stress-tested against *impersonation*. Accordingly, this chapter investigates a **SRQ3**: *How effective are transformer-based classifiers for phishing text-based attacks, and to what extent do they generalise to impersonation-driven phishing text-based attacks when only email body content is available?*

We address this by proposing a transformer-based detector that analyses email body text rather than metadata. Our model combines transformer based model with a Bidirectional Long Short-Term Memory (BiLSTM) layer to capture important sequential text relationships.

The following sections detail the related work, proposed model, experiment, and evaluation of the proposed model, demonstrating its effectiveness in detecting phishing text-based attacks while addressing key challenges in text-based deception detection.

[0]This chapter is based on the publication: Almutairi, A. M., Kang, B., & Fadhel, N. (2023). *The Effectiveness of Transformer-Based Models for BEC Attack Detection*. In: Li, S., Manulis, M., Miyaji, A. (eds) *Network and System Security*. NSS 2023. Lecture Notes in Computer Science, vol 13983. Springer, Cham. https://doi.org/10.1007/978-3-031-39828-5_5

7.2 Related Work

A detailed survey appears in Chapter 3, shows that existing phishing text-based attacks detection systems generally fall into three categories, each with notable limitations. First, many approaches depend heavily on metadata signals—such as SPF/DKIM validation, sender IP, or domain reputation—which become ineffective when an attacker compromises a legitimate mailbox. Second, several models rely on hand-engineered lexical or stylistic features that adversaries can easily obfuscate. Third, some deep learning methods apply aggressive pre-processing (e.g., lowercasing, stemming, punctuation removal), stripping out the subtle textual cues often exploited in BEC attacks. Unlike prior work, the proposed method retains punctuation and casing, allowing it to capture syntactic and stylistic signals critical for early-stage phishing text-based attacks detection for example BEC attack, especially in impersonation scenarios where metadata appears legitimate.

7.3 Proposed Model

Rationale. Transformers (e.g., BERT) capture rich contextual semantics, yet short, formulaic business emails may benefit from additional sequence modelling to retain stylistic rhythm. We therefore augment a compact encoder (DistilBERT) with a BiLSTM layer.

Architecture. The pipeline (Fig. 7.1) comprises: (i) DistilBERT for contextual embeddings, (ii) a BiLSTM for bidirectional sequence dynamics, (iii) a feed-forward classifier with softmax.

TABLE 7.1: Hyperparameters (as tuned on validation).

Hyperparameter	Value
Max token length	256
Batch size	16
Learning rate	2×10^{-5} (grid: $[1,3] \times 10^{-5}$)
Epochs	3
LSTM hidden size	50

7.4 Experiments

7.4.1 Datasets and Splits

Two public corpora were used for comparability/reproducibility, and a stress-test corpus for BEC mimicry:

- **Fraud Email Detection:** A benchmark dataset comprising 5,187 phishing and 6,742 legitimate messages, introduced by Radev (2008).

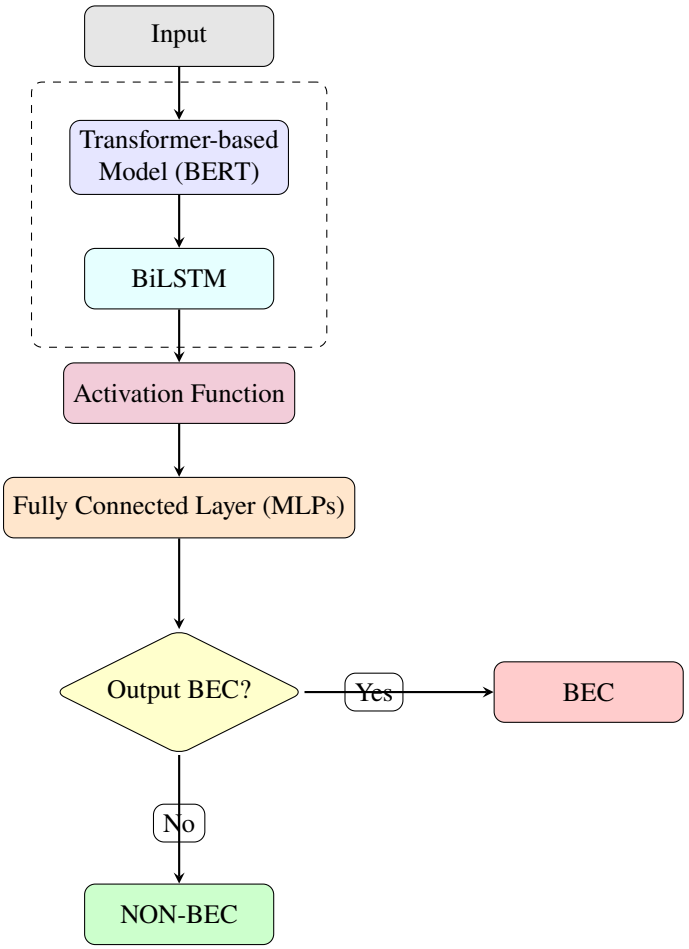


FIGURE 7.1: DistilBERT+BiLSTM model workflow.

- **TREC 2007:** A widely used collection consisting of 50,199 phishing and 25,220 legitimate emails, originally presented by Macdonald et al. (2007) for the TREC Spam Track.
- **Impersonation BEC (synthetic):** A constructed dataset featuring paired legitimate and impersonation-style BEC messages per author, as detailed in chapter 6. It includes five Enron authors and follows an author-disjoint train/test split to preserve authorship integrity.

Split policy. This experiment follows the *Common Experimental Setup*. (Chapter 4): 70% train, 10% validation, 20% test, stratified by label. For impersonation, authors are disjoint across splits to avoid overfitting to idiosyncratic style.

Text-only constraint. To isolate linguistic signal, Headers/metadata, attachments, raw URLs/HTML tags, and boilerplate signatures are removed during preprocessing.

7.4.2 Training Protocol and Metrics

The unified experimental settings from Chapter 4 are adopted: Python 3.x, PyTorch, and HuggingFace Transformers; cross-entropy loss; the AdamW optimizer; early stopping based on

validation F1-score; and evaluation metrics including Accuracy, Precision, Recall, F1-score, and macro-averaged ROC–AUC.

7.4.3 Baselines

Two classical text baselines use TF–IDF features:

- **Random Forest** (bag-of-words TF–IDF).
- **XGBoost** (bag-of-words TF–IDF).

Preprocessing for baselines follows standard practice (lowercasing, tokenization; punctuation/numbers/stopwords removed) to match prior work.

7.5 Results

7.5.1 Results on Public Phishing Corpora

Table 7.2 shows that DistilBERT+BiLSTM outperforms the baselines on both corpora.

TABLE 7.2: Performance on Fraud and TREC07 (best in bold).

Model	**Fraud**				**TREC07**			
	Prec	**Rec**	**F1**	**Acc**	**Prec**	**Rec**	**F1**	**Acc**
DistilBERT+BiLSTM	**99.26**	**99.41**	**99.33**	**99.25**	**99.19**	**99.21**	**99.20**	**99.21**
Random Forest	98.34	96.37	97.34	97.02	98.86	98.87	98.86	98.87
XGBoost	95.64	99.18	97.38	96.98	98.74	98.69	98.71	98.73

7.5.2 Comparison to recent Studies.

We also compare with a recent BERT+CNN+BiGRU pipeline on TREC07 and with a BiLSTM-Attention pipeline on Fraud (Tables 7.3–7.4). Our model is competitive or superior while remaining purely content-based.

TABLE 7.3: Comparison on **TREC07**.

Reference	**Method**	**Acc (%)**	**Prec (%)**	**Rec (%)**	**F1 (%)**
Alguliyev et al. (2024)	BERT + ConvNet + BiGRU	98.67	98.79	98.39	98.59
This work	DistilBERT + BiLSTM	**99.21**	**99.19**	**99.21**	**99.20**

TABLE 7.4: Comparison on **Fraud**.

Reference	Method	Acc (%)	Prec (%)	Rec (%)	F1 (%)
Xiao and Jiang (2020)	BiLSTM-Attention	91.51	91.75	91.49	91.58
This work	DistilBERT + BiLSTM	**99.25**	**99.26**	**99.41**	**99.33**

7.5.3 Replication under identical preprocessing.

The BERT+BiGRU+CNN model from Alguliyev et al. (2024) was replicated using the original preprocessing steps and hyperparameter settings. As shown in Table 7.5, DistilBERT+BiLSTM remains competitive in comparison.

TABLE 7.5: Side-by-side replication on Fraud and TREC07.

Model	Fraud				TREC07			
	Prec	**Rec**	**F1**	**Acc**	**Prec**	**Rec**	**F1**	**Acc**
Replicated Alguliyev et al. (2024)	99.23	99.13	99.18	99.19	97.32	95.84	96.52	96.93
(This work)	**99.26**	**99.41**	**99.33**	**99.25**	**99.19**	**99.21**	**99.20**	**99.21**

7.5.4 Results on Impersonation-Based BEC

Dataset. We construct a style-mimicry corpus (chapter 6) where BEC messages imitate the tone and phrasing of specific Enron authors. Train/test splits are author-disjoint.

TABLE 7.6: Classification on impersonation-based BEC (author-disjoint).

Metric	Average	AUC
Precision	68.92	
Recall	65.26	58.09
F1-Score	63.49	
Accuracy	65.26	

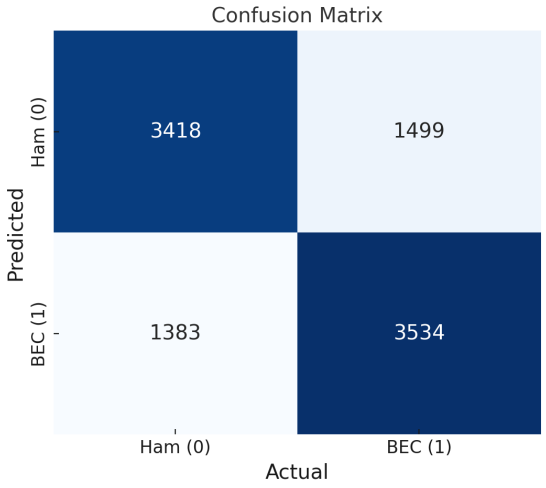


FIGURE 7.2: Confusion matrix on impersonation-based BEC emails.

7.6 Discussion and Analysis

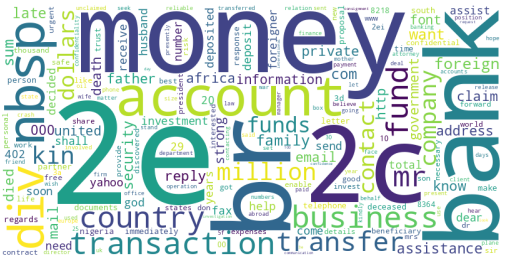
Chapter Contribution

This chapter contributes two fold: (1) It establishes an efficient, reproducible baseline for transformer-based BEC detection that operates solely on email body content, achieving strong performance across public datasets without relying on metadata or handcrafted features. (2) It presents a principled *negative result* under impersonation-style attacks, demonstrating the limitations of content-only approaches when semantic cues are deliberately camouflaged within legitimate writing styles. Unlike many prior phishing detection approaches, which either depend on metadata (e.g., SPF/DKIM) or handcrafted lexical features, this model operates exclusively on raw email text. Its hybrid architecture—combining contextual encoding via BERT with sequential modelling via BiLSTM—offers an efficient alternative to more complex CNN+GRU architectures, while achieving competitive or superior results on benchmark datasets (Tables 7.3–7.5).

7.6.1 Linguistic Feature Analysis

To explore the lexical patterns learned by the model, an analysis was conducted on discriminative terms within two benchmark corpora: the Fraud dataset and TREC07. Word clouds were generated for each using a TF–IDF weighting scheme, with tokens filtered through ANOVA F-statistics against class labels to identify the top 300 most informative terms.

Key Observations. Figure 7.3 presents word clouds of the most informative lexical cues in the Fraud and TREC07 datasets, based on their discriminative power in email classification tasks.



(A) Fraud dataset: top discriminative terms.

(B) TREC07 dataset: top discriminative terms.

FIGURE 7.3: Word clouds of the most informative terms in the Fraud and TREC07 corpora.

- **Fraud dataset:** Lexical indicators are dominated by financial terminology (*payment*, *money*, *transaction*) and formal address cues (*dear*, *please*, *Mr.*), suggesting attempts to emulate legitimate professional tone while delivering fraudulent intent.
- **TREC07 dataset:** Discriminative tokens are primarily link-related (*www*, *https*, *org*), reflecting phishing's heavy reliance on embedded URLs and external redirection mechanisms.

These lexical patterns reinforce prior findings: traditional phishing detection often hinges on shallow surface cues, whereas BEC impersonation attacks are designed to blend into legitimate correspondence through stylistic mimicry. This supports the transition toward author-style modeling, which is explored in Chapters 8 and 9.

7.6.2 Limitations and Implications

Although the proposed model demonstrated strong performance on benchmark phishing datasets, its accuracy dropped to 65% when evaluated on the synthetic impersonation corpus. This moderate performance highlights inherent challenges in detecting Business Email Compromise (BEC) when only the email body content is available, particularly under impersonation scenarios that lack overt anomalies.

These results suggest that traditional content-based phishing detectors are ill-suited for impersonation-style BEC attacks, where malicious intent is deliberately concealed within legitimate stylistic patterns. The observed 65% accuracy does not reflect a flaw in the model architecture, but rather the intrinsic difficulty of the task—especially when deceptive content mimics the tone, structure, and vocabulary of the impersonated sender. This performance boundary reinforces the need for more identity-sensitive modelling approaches, which go beyond semantic detection to capture personalized stylistic signals.

These findings highlight a fundamental limitation: conventional phishing detectors, which rely primarily on lexical and semantic signals, are insufficient for handling impersonation-driven BEC attacks. This motivates the shift toward incorporating authorship-aware verification techniques,

as detailed in Chapter 8, and their integration within a unified multi-task learning framework in Chapter 9.

7.7 Chapter Summary

This chapter introduced a transformer-based detector that combines DistilBERT embeddings with a BiLSTM layer for sequential modelling. The model was evaluated on two widely used public phishing corpora (Fraud and TREC07), where it consistently outperformed traditional baselines and recent neural approaches, achieving state-of-the-art performance. Word-cloud analysis confirmed that high accuracy on these corpora is largely driven by the presence of surface lexical cues such as URLs, financial terms, and politeness markers.

However, when evaluated against a synthetic impersonation-based BEC dataset designed to mimic genuine writing styles, the model's performance dropped markedly to 65% accuracy with an AUC of 0.58. This highlights a key limitation: phishing datasets, while useful for benchmarking, do not capture the linguistic realism of BEC attacks, where attackers impersonate trusted insiders using plausible tone and style. As a result, content-only models trained on generic phishing data cannot be relied upon to detect sophisticated impersonation attempts.

The findings therefore serve two purposes. First, they demonstrate that transformer-based models are effective at phishing detection when surface cues are present. Second, and more importantly, they expose the insufficiency of such models for detecting BEC, thereby justifying the need for additional mechanisms that verify authorship consistency. This observation directly motivates the next chapter, which introduces **BiBERT-AV**, a Siamese-style authorship verification model designed to capture stable stylistic signatures of legitimate users and detect identity-spoofing BEC attacks.

Chapter 8

BiBERT-AV: A Siamese Network for Authorship Verification

8.1 Introduction

Chapter 7 showed that content-only transformers can reach state-of-the-art performance on public phishing corpora, yet degrade markedly under *impersonation*, which is central to Business Email Compromise (BEC). This chapter addresses that gap with an *authorship verification* (AV) module designed to check whether an email's writing style is consistent with the claimed sender.

We present **BiBERT-AV**, a Siamese architecture that combines transformer embeddings with sequence modelling to capture both semantic context and stylistic rhythm. The chapter answers **SRQ4**: *How do transformer-based Siamese networks perform in authorship verification of business emails compared to traditional stylometric and recent neural methods?*

8.2 Related Work

This chapter builds on the broader body of AV research reviewed in Chapter 3, which traces the evolution of AV from handcrafted stylometry to deep, transformer-based representations. Chapter 3 (Section 3.3) provides a comprehensive overview of AV techniques, including traditional feature-based models, hybrid methods, and modern Siamese architectures tailored to short-text domains such as emails.

The present model, BiBERT-AV, follows this trajectory by adopting a lightweight Siamese framework with a shared encoder and a learned similarity head. Unlike prior cosine-only or

[0]This chapter is based on the publication: Almutairi, A. M., Kang, B., & Al Hashimy, N. (2023). *BiBERT-AV: Enhancing Authorship Verification Through Siamese Networks with Pre-trained BERT and Bi-LSTM*. In: Manulis, M., Miyaji, A., Zhang, Y. (eds) *International Conference on Ubiquitous Security*. Lecture Notes in Computer Science, vol 13984. Springer, Cham. https://doi.org/10.1007/978-3-031-xxxxx-x

contrastive-loss approaches, BiBERT-AV is designed for real-time, mimic-resistant AV in Business Email Compromise (BEC) contexts, and is evaluated under author-disjoint, open-set conditions to simulate realistic enterprise deployment scenarios. For detailed comparisons to baseline AV methods, including task-specific variants and Enron-focused studies, see Table 3.4 in Chapter 3.

8.3 Model: BiBERT-AV

8.3.1 Architecture

Figure 8.1 illustrates BiBERT-AV, a Siamese neural network designed to verify whether two email messages were authored by the same individual. The model integrates transformer-based contextual encoders with sequential pattern extraction, enabling it to detect stylistic consistency between messages beyond superficial word overlap.

Each input email is processed through a shared BERT encoder, which produces contextual token embeddings. To enhance sensitivity to word order and punctuation patterns, these embeddings are further refined using a Bidirectional LSTM (BiLSTM) layer. The final representations capture both semantic content and syntactic style.

During training, the model receives pairs of email bodies labelled as either same-author or different-author. The resulting embeddings are passed through dense layers and combined to produce a similarity score, which is then mapped to a binary classification.

At test time, an incoming email is encoded and compared against a precomputed reference embedding for the claimed sender. These reference vectors are created in advance by averaging the encoder outputs from multiple known emails written by that sender.

8.3.2 Training Objective

Given label $y \in \{0, 1\}$ (same/different author), we minimise binary cross-entropy:

$$\mathcal{L} = -\big[y \log p + (1 - y) \log(1 - p)\big].$$

We report Accuracy, Precision, Recall, F1, and macro ROC–AUC. At inference, a *reference embedding* per author is computed as the mean of that author's known emails; an incoming email is compared against the claimed author's reference.

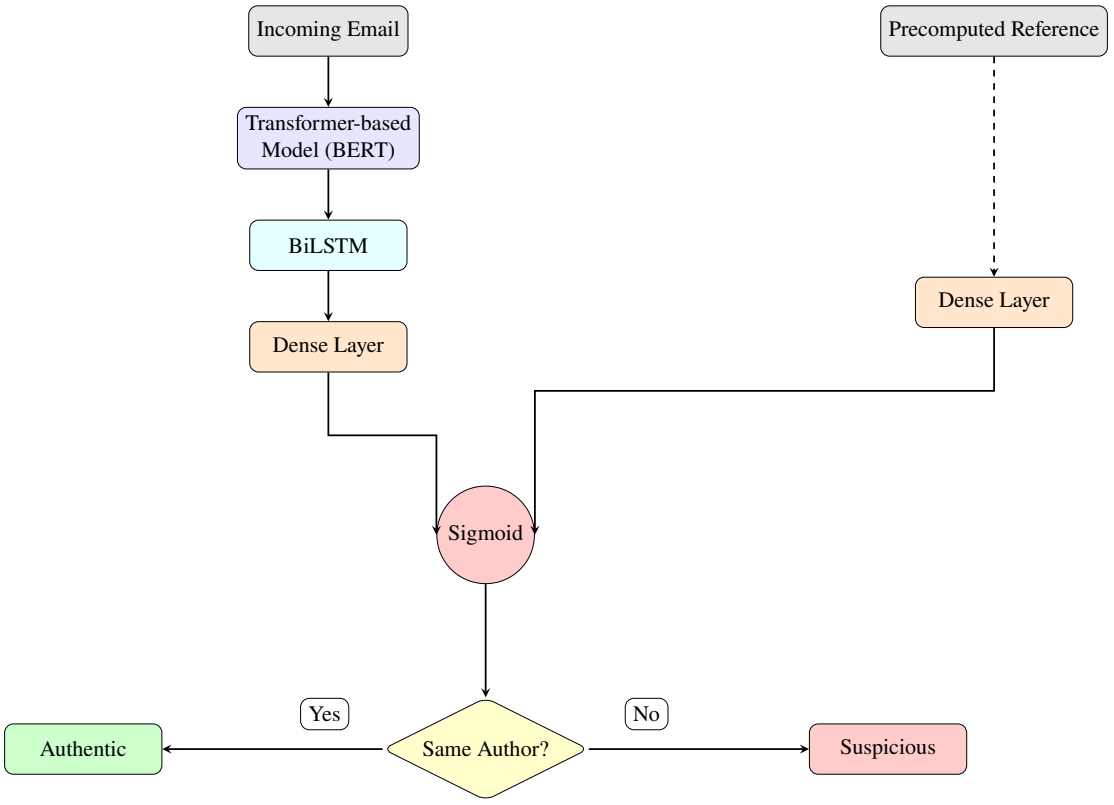


FIGURE 8.1: BiBERT-AV architecture: comparing the incoming email to a precomputed reference embedding of the claimed author.

8.4 Datasets and Splits

8.4.1 Enron Email for AV

The Enron corpus was used as the primary dataset, and it is a well-known business email dataset. All metadata, headers, forwarded content, and attachments were stripped, retaining only the cleaned email body text.

Emails were grouped by sender to generate labelled pairs:

- **Positive pairs:** Two emails written by the same sender.
- **Negative pairs:** Emails written by different senders.

We evaluated the model on author subsets of increasing size—2, 5, 10, 20, and 50 authors—selected based on the volume of emails per sender. For each subset, the data was split using stratified sampling as described in Section 4.5.3.

At inference, the incoming email is encoded into a vector $\mathbf{h}_a$, which is then compared against a precomputed reference vector $\mathbf{h}_b$ for the claimed sender. The reference vector is generated

by averaging the encoder outputs of that sender's emails during training. This avoids repeated computation and simulates deployment conditions where historical embeddings are pre-stored.

8.4.2 Mimic Dataset for Impersonation

This evaluation uses the *Authorship Mimicry Dataset* described in Chapter 6, which contains synthetic mimic Enron emails generated to imitate the writing styles of five Enron authors. We focus exclusively on this mimicry subset because the objective here is authorship verification under stylistic impersonation, independent of semantic content related to BEC.

8.5 Hyperparameters and Rationale

Table 8.1 lists all hyperparameters along with their empirical or theoretical rationale. Most values were selected based on a combination of validation set performance, ablation studies, and practical deployment considerations (e.g., latency and memory constraints). This ensures both accuracy and feasibility in real-world enterprise environments.

TABLE 8.1: Hyperparameters used in BiBERT-AV training and their justifications.

Parameter	Value	Justification
Maximum input length	256 tokens	Covers majority of business emails without truncation
Learning rate	3×10^{-5}	Best performance in grid search $[1e\text{–}5, 5e\text{–}5]$
Loss function	Binary Cross-Entropy	Suited for binary similarity classification
Activation	Sigmoid	Outputs probability for binary decision
Epochs	10	Converged without overfitting on validation set
Batch size	16	Balances memory constraints and convergence speed

8.6 Results

8.6.1 Results on Enron Email Dataset

Table 8.2 shows the model's performance across different author pool sizes. BiBERT-AV maintains high precision and recall as the number of candidate authors increases, demonstrating robustness to growing verification complexity.

8.6.2 Comparison with Existing Methods

We compared BiBERT-AV against the Siamese BERT model from Tyo et al. (2021), using identical data splits and metrics. Table 8.3 shows that BiBERT-AV consistently outperforms the baseline across all author subsets.

TABLE 8.2: Authorship verification performance across author pool sizes.

Authors	Precision	Recall	F1-score	Accuracy
Two	99.00	99.00	99.00	99.00
Five	98.00	98.00	98.00	98.00
Ten	98.00	98.00	98.00	98.00
Twenty	95.00	95.00	95.00	95.00
Fifty	90.00	93.00	90.00	90.00

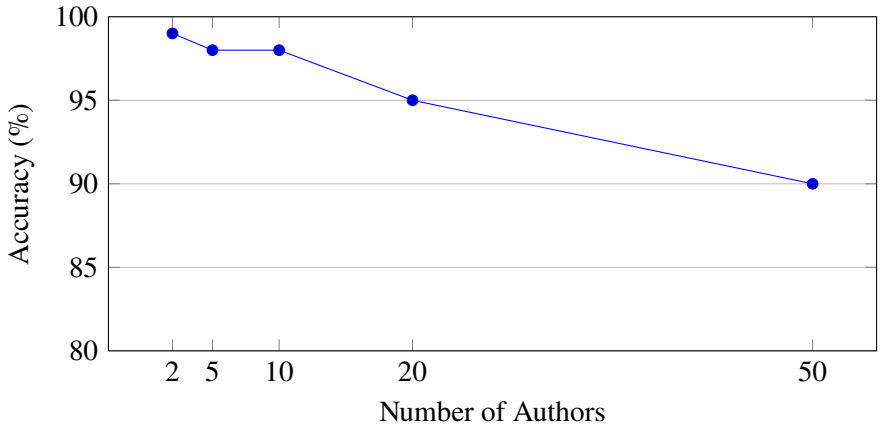


FIGURE 8.2: Accuracy vs. author-pool size on Enron.

TABLE 8.3: Comparison of BiBERT-AV and Siamese BERT on Enron dataset.

Authors	Siamese BERT Tyo et al. (2021)				BiBERT-AV			
	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score	Accuracy
Two	68.00	87.00	77.00	77.00	99.00	99.00	99.00	99.00
Five	77.00	71.00	74.00	79.00	98.00	98.00	98.00	98.00
Ten	83.00	76.00	79.00	80.00	98.00	98.00	98.00	98.00
Twenty	75.00	73.00	74.00	74.00	95.00	95.00	95.00	95.00
Fifty	49.00	81.00	61.00	50.00	90.00	93.00	90.00	90.00

8.6.3 Authorship Mimicry Dataset Evaluation

Using the style-mimicry subset described in Chapter 6 (Section 6.2.6.2), BiBERT-AV demonstrates strong performance in detecting impersonation-based emails. The model maintains high confidence even when adversarial samples closely emulate the writing style of legitimate authors, underscoring its effectiveness in challenging mimicry scenarios.

TABLE 8.4: BiBERT-AV on the Authorship Mimicry Dataset.

Metric	**Macro Avg**	**AUC**
Precision	96.10%	
Recall	95.51%	98.97%
F1-score	95.80%	
Accuracy	95.82%	

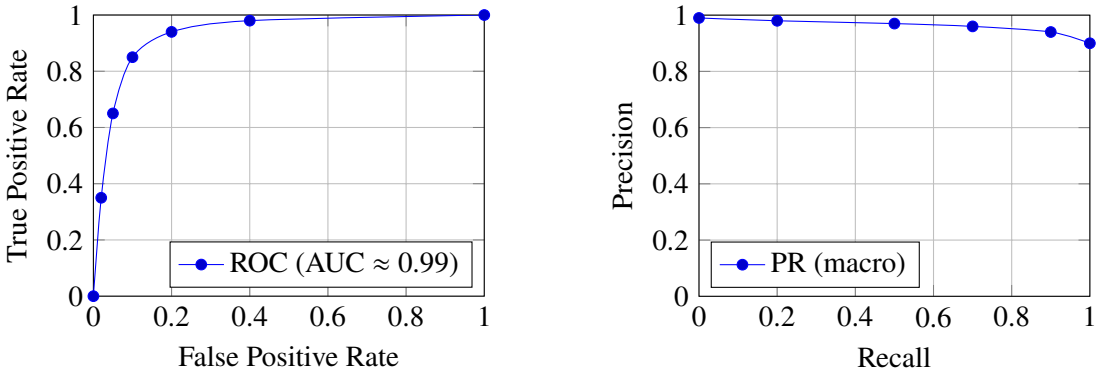


FIGURE 8.3: Mimicry subset: ROC and Precision–Recall curves. BiBERT-AV retains high discriminative power under style-consistent deception.

8.7 Discussion

Contribution. This chapter introduced **BiBERT-AV**, a mimic-resistant AV model tailored for enterprise email security. Operating under strict author-disjoint settings, BiBERT-AV models individual writing style through a supervised Siamese architecture with sequence pooling and a learned similarity function. Rather than asking whether an email looks malicious, it asks: *"Does this message sound like it was written by the claimed sender?"* This stylistic perspective provides an orthogonal defence to traditional phishing detectors, especially in scenarios involving internal impersonation and Business Email Compromise (BEC). Compared to prior AV approaches such as the Siamese BERT model proposed by Tyo et al. (2021), BiBERT-AV offers a more robust treatment of stylistic similarity through sequence-level pooling and a learned similarity function, rather than relying on fixed-distance metrics. Furthermore, unlike unsupervised clustering or metadata-dependent AV systems discussed in Chapter 3, our model operates in a fully supervised, author-disjoint regime and is explicitly evaluated under mimicry conditions. This design enables BiBERT-AV to resist impersonation attacks and generalise across unseen authors—two critical gaps unaddressed by most prior AV work.

Performance and Robustness. As shown in Table 8.4, BiBERT-AV consistently outperforms cosine-only baselines across both standard and impersonation-focused evaluations. Its performance remains robust even as the author pool expands—an essential feature for real-world deployment across large organisations. Figure 8.3 illustrates this reliability under mimicry conditions: the

model sustains high precision–recall and ROC performance despite semantic ambiguity and lexical overlap introduced by stylistic deception.

Figure 8.4 provides further insight into the model's behaviour. The left panel shows that BiBERT-AV exhibits well-calibrated predictions, aligning predicted probabilities with empirical accuracy—an important property for operational decision-making. The right panel visualises a 2D t-SNE projection of email embeddings, where clusters show clean separation between authors even under mimicry conditions, reflecting the model's ability to learn stylistically meaningful representations.

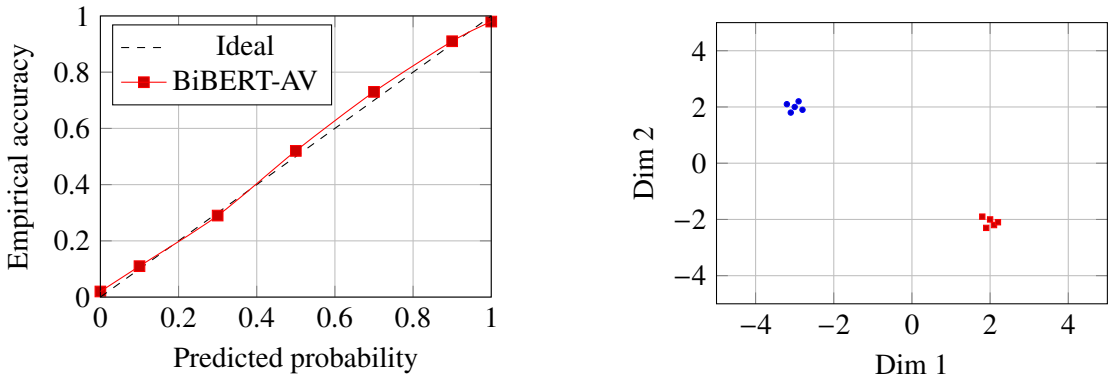


FIGURE 8.4: Left: Reliability curve showing calibration quality. Right: Stylometric clusters in 2D projection, suggesting author separation under mimicry.

Implications. Authorship Verification, when treated as a supervised classification problem with learned embeddings, offers a robust and scalable alternative. While AV is not a standalone defence, it plays a critical role in layered email security architectures by reintroducing identity verification through linguistic style—an attribute difficult to forge without long-term access or behavioural leakage.

8.8 Chapter Summary

BiBERT-AV, a Siamese transformer+BiLSTM with a learned similarity head, verifies author identity from email body text alone. It maintains high performance across growing author pools and detects style-mimicry emails drawn from the chapter 6 dataset with strong precision/recall and near-perfect AUC. The model supplies the *authorship layer* needed to complement content detectors in BEC defence. The next chapter unifies these signals with MTL for end-to-end BEC detection under impersonation.

Chapter 9

A Multi-Task Learning Framework for Joint BEC Detection and Authorship Verification

9.1 Introduction

Previous chapters presented two independent models addressing distinct aspects of Business Email Compromise (BEC) detection. Chapter 7 introduced a transformer-based classifier to identify semantic anomalies and deceptive intent in email content, whereas Chapter 8 presented BiBERT-AV, a model that verifies authorship by analyzing writing style.

However, real-world BEC attacks often blend semantically plausible lures with stylistic mimicry to evade single-axis detectors. As detailed in Chapter 5, attackers may craft messages that read as legitimate yet subtly deviate from an executive's usual writing style, or they may spoof a trusted sender's style while embedding malicious intent. Traditional tools relying on either content or metadata frequently fail under these hybrid tactics, especially when metadata is missing or compromised.

To address this gap, we propose a Multi-Task Learning (MTL) architecture that jointly models semantic deception and authorial consistency directly from email text. This framework targets SRQ5: *How does integrating BEC detection and authorship verification into a single system affect overall accuracy and operational cost?* By sharing a common encoder and employing task-specific heads, our approach enhances detection effectiveness and reduces inference overhead. This dual-task design is particularly critical for high-value targets—such as executives—whose communications require both semantic scrutiny and authorial validation to prevent sophisticated impersonation attacks.

[0]This chapter is based on the manuscript: Almutairi, A., Kang, B., and Al Hashimy, N. (2024). *Integrating Business Email Compromise Detection and Authorship Verification Through Multi-Task Learning*. Submitted and currently under review at the *Journal of Information Security and Applications*.

Background and Literature Context

This chapter builds upon the Multi-Task Learning (MTL) literature reviewed in Chapter 3, specifically Section 3.4, which surveyed applications of MTL in NLP and deception detection. Prior work has shown that related tasks—such as sentiment, novelty, or emotion classification—can enhance robustness and generalization when jointly modeled.

Informed by these findings, this chapter introduces a unified MTL framework tailored to the hybrid nature of Business Email Compromise (BEC), where semantic deception and stylistic impersonation often co-occur. By combining BEC detection with authorship verification in a shared encoder setting, the proposed model leverages cross-task signals to improve resilience against subtle, impersonation-driven attacks.

9.2 Proposed Framework

This section introduces a unified Multi-Task Learning (MTL) framework that jointly addresses Business Email Compromise (BEC) detection and AV. The model is designed to enhance the detection of BEC fraud in the early stage by learning both semantic and stylistic patterns from the text email body.

As established in **Chapters 7** and **8**, BEC detection and AV address distinct but complementary objectives. BEC detection identifies indicators of malicious intent, while AV determines whether a message is stylistically consistent with the claimed sender. Since real-world BEC attacks often exhibit plausible content but deviate from an author's usual writing style, combining these two capabilities can enhance detection even when emails are crafted to appear legitimate.

However, integrating BEC and AV into a single model presents several design challenges:

- **Different Task Requirements:** BEC detection and AV focus on different types of signals—semantic content versus writing style. Using the same model layers for both without separation can weaken their individual performance.
- **Mismatch in Output Structure:** BEC detection predicts a single label for each email (malicious or not), whereas AV compares two emails and predicts whether they come from the same author.
- **Training Conflicts:** The two tasks use different loss functions (classification vs. contrastive), so training them together requires careful balancing to avoid one task dominating the learning process.

To address these challenges, the proposed MTL framework adopts the following design:

- **Shared Encoder:** A BERT–BiLSTM encoder that encodes both contextual and stylistic features from input text.
- **Task-Specific Heads:**
  - A classification head for BEC detection.
  - A Siamese-style contrastive head for AV.
- **Joint Optimization:** The total loss combines both task objectives:

$$\mathcal{L}_{\text{total}} = \alpha \mathcal{L}_{\text{BEC}} + \beta \mathcal{L}_{\text{AV}}, \tag{9.1}$$

  where $\alpha = \beta = 1$ in this thesis.

This structure supports efficient learning by sharing a common encoder while preserving specialization through task-specific heads and loss functions. The result is a content-driven detection system capable of identifying BEC attacks even when metadata is unavailable or manipulated.

The following section outlines the architecture and training methodology used to implement and evaluate the framework.

9.2.1 Methodology

This section details the implementation of the proposed Multi-Task Learning (MTL) framework, which jointly performs Business Email Compromise (BEC) detection and AV using a shared neural architecture.

9.2.1.1 Framework Architecture

As illustrated in Figure 9.1, the framework consists of a shared encoder and two task-specific heads. The shared encoder integrates:

- **BERT:** A transformer pre trained on general-domain corpora, used to extract contextual embeddings from email body text.
- **BiLSTM:** A bidirectional LSTM layer applied to the transformer output to encode sequential and stylistic patterns.

This encoder is trained under a hard parameter-sharing regime, meaning both tasks update the same parameters during backpropagation. This setup promotes inductive transfer while reducing model complexity.

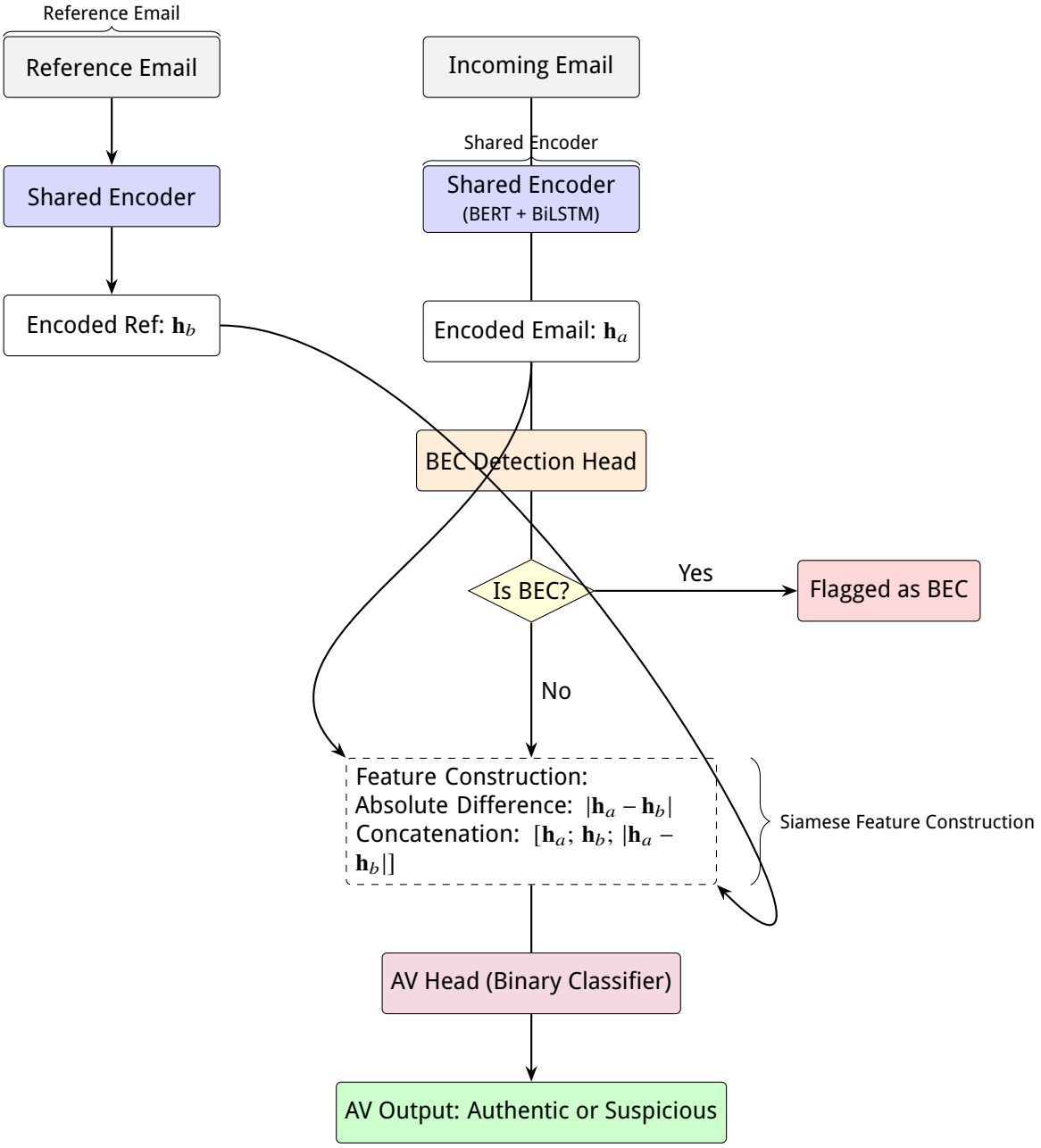


FIGURE 9.1: MTL inference pipeline for joint BEC detection and authorship verification. A shared encoder generates embeddings for both the incoming and reference emails. BEC is first classified directly; if not flagged, authorship verification compares stylistic features to detect impersonation.

Task-Specific Heads.

- **BEC Detection Head (see Chapter 7):** A fully connected layer with sigmoid activation, responsible for classifying whether an email exhibits BEC-related characteristics.
- **AV Head (see Chapter 8):** A contrastive Siamese classifier that compares the embedding of an incoming email with a reference embedding derived from the claimed author's historical messages.

Training objectives. Each task-specific head is optimized with a binary cross-entropy–with–logits loss.

$$\mathcal{L}_{\text{BEC}} = \frac{1}{N}\sum_{i=1}^{N}\Big[\max\big(z_i^{\text{BEC}},0\big) - y_i^{\text{BEC}}\, z_i^{\text{BEC}} + \log\big(1+e^{-|z_i^{\text{BEC}}|}\big)\Big], \tag{9.2}$$

$$\mathcal{L}_{\text{AV}} = \frac{1}{N}\sum_{i=1}^{N}\Big[\max\big(z_i^{\text{AV}},0\big) - y_i^{\text{AV}}\, z_i^{\text{AV}} + \log\big(1+e^{-|z_i^{\text{AV}}|}\big)\Big]. \tag{9.3}$$

In (9.2)–(9.3), $z \in \mathbb{R}$ is the raw logit (pre-sigmoid) and $y \in \{0,1\}$ is the label. This is algebraically equivalent to binary cross-entropy on the sigmoid probability,

$$\mathcal{L}_t \;=\; -\frac{1}{N}\sum_{i=1}^{N}\Big(y_i^t \log\sigma(z_i^t) + (1-y_i^t)\log\big(1-\sigma(z_i^t)\big)\Big), \quad \sigma(z)=\frac{1}{1+e^{-z}}, \;\; t\in\{\text{BEC},\text{AV}\}, \tag{9.4}$$

Total joint loss.

$$\mathcal{L}_{\text{total}} = \alpha\,\mathcal{L}_{\text{BEC}} + \beta\,\mathcal{L}_{\text{AV}}, \qquad \alpha=\beta=1. \tag{9.5}$$

The AV head operates on a pair of inputs: the encoded representation of the incoming email ($\mathbf{h}_a$) and a reference embedding of the claimed author ($\mathbf{h}_b$). These embeddings are combined using element-wise absolute difference and concatenation, then passed to a fully connected layer for binary classification.

This architecture supports content-only verification, allowing the model to detect BEC threats and validate authorship even in the absence of metadata or headers. The next section outlines the training configuration, dataset construction, and evaluation procedures.

9.2.2 Training and Optimization Strategy

This section outlines the training procedure for the proposed Multi-Task Learning (MTL) framework. The model is trained end-to-end using a joint loss function that combines objectives for both Business Email Compromise (BEC) detection and AV. The inference logic is described in Algorithm 1, and the overall training procedure is summarized in Table 9.1.

9.2.3 Dataset Construction and Preprocessing

A composite dataset was constructed to jointly support the BEC and AV tasks, combining real and synthetic samples. AV instances were generated using mimicry prompts targeting authorial

Algorithm 1 MTL Email Security Framework (Inference Logic)

Require: **New Email**, **Reference Embedding from Historical Emails**
Ensure: Classification result
Encode the new email using the shared encoder to obtain $\mathbf{h}_a$
Compute BEC prediction
if Predicted as BEC **then**
    Flag email as malicious
else
    Retrieve the reference embedding $\mathbf{h}_b$ of the claimed sender
    Compute absolute difference: $\mathbf{d} = |\mathbf{h}_a - \mathbf{h}_b|$
    Concatenate feature vectors and compute AV prediction
    if AV score $\geq \theta$ **then**
        Output: Authentic
    else
        Output: Suspicious; escalate for review
    end if
end if

TABLE 9.1: Training hyperparameters.

Parameter	Value
Batch size	16
Optimizer	AdamW
Learning rate	2×10^{-5}
Epochs	10 (early stop)
Cross-validation	5-fold
Early stopping patience	2
Hidden size (BiLSTM)	128
Dropout rate (heads)	0.1
Loss weights	α=1.0, β=1.0
Random seed	42

Notes: Learning rate was grid-searched over $[1, 3] \times 10^{-5}$; 2×10^{-5} was selected at the development-set plateau. Equal loss weighting was chosen after a sweep over $\{(0.5, 0.5), (0.6, 0.4), (0.7, 0.3)\}$ showed negligible macro-F1 differences and better stability across seeds.

style, while BEC samples represent a range of phishing and impersonation attacks. The complete construction pipeline—including seed selection, prompt templates, quality-control filters, and ethical safeguards—is documented in chapter 6.

9.2.3.1 Dataset Composition

The multi-task training dataset was constructed to support both Business Email Compromise (BEC) detection and AV, combining real and synthetic samples to simulate realistic impersonation scenarios; full construction details are provided in Chapter 6. It consists of two task-specific components:

1. **Authorship Verification (AV) Dataset:**
    - **Real Emails:** 5,000 messages authored by five high-volume individuals from the Enron corpus, selected based on availability and volume.
    - **Synthetic Emails:** 5,000 LLaMA-generated emails fine-tuned to imitate the writing style of each target author.
    - **Pair Construction:** Email pairs were created to support contrastive training for binary authorship verification:
        - *Same-author pairs:* Two real emails written by the same Enron author.
        - *Different-author pairs:* Pairs consisting of either emails from two distinct authors, or a real email paired with a synthetic mimic.
2. **Business Email Compromise (BEC) Dataset:**
    - **Real BEC Emails:** 21 samples sourced from public disclosures and academic archives.
    - **Synthetic BEC:** 1,050 LLaMA-generated BEC-style emails.
    - **Additional Phishing Corpora:** CEAS08, TREC07, LingSpam, and SpamAssassin.

To ensure robust evaluation and prevent information leakage:

- **AV splits were based on email instances**, allowing each author's writing style to be learned from historical emails and tested on unseen samples by the same author.
- **BEC and non-BEC samples were stratified** to maintain class balance across training, validation, and test sets.

This dataset design enables the model to jointly learn semantic deception cues (for BEC detection) and stylistic consistency patterns (for AV), while supporting scalable and realistic evaluation in both tasks.

9.3 Baseline Model

To assess the added value of joint training in the proposed Multi-Task Learning (MTL) framework, we compare it against a sequential Transfer Learning (TL) baseline. This baseline preserves the same core architecture and training schedule but treats the tasks independently rather than concurrently.

The TL baseline involves the following two-stage process:

- **Stage 1 – BEC Task:** A shared encoder comprising BERT followed by a BiLSTM layer is trained solely on the BEC classification task.

- **Stage 2 – AV Task:** The BEC classification head is replaced with a Siamese-style contrastive head for authorship verification. The model is then fine-tuned on the AV dataset using a reduced learning rate.

To mitigate catastrophic forgetting during the second phase, fine-tuning was performed with early stopping based on validation performance. All other variables—model hyperparameters, data splits, and preprocessing—were held constant across both the TL and MTL settings. Each model was trained using five different random seeds, and the final reported results represent the average across these runs. This controlled setup ensures that any observed differences in performance are attributable solely to the training paradigm (i.e., joint versus sequential learning).

9.4 Classification performance

TABLE 9.2: Performance on the validation (*eval*) and held-out (*test*) sets. Best scores per column appear in bold.

		Eval				**Test**			
Model	**Task**	Acc.	Prec.	Rec.	F_1	Acc.	Prec.	Rec.	F_1
TL (sequential)	BEC	0.86	0.89	0.85	0.86	0.85	0.78	0.71	0.82
	AV	0.92	0.91	0.90	0.91	0.91	0.90	0.91	0.91
MTL (joint)	BEC	**0.98**	**0.97**	**0.97**	**0.97**	**0.98**	**0.96**	**0.96**	**0.97**
	AV	**0.94**	**0.93**	**0.95**	**0.94**	**0.93**	**0.92**	**0.93**	**0.93**

These findings remained stable across five independent training runs with different random seeds, showing a standard deviation of less than 1.5% across all metrics. Notably, the greatest performance gains were observed in recall—an essential metric in security systems, where failing to detect malicious activity (false negatives) can have severe consequences.

9.4.1 Receiver–operating characteristics (ROC)

The ROC plot in Figure 9.2 shows that the MTL curve (orange) consistently sits above the TL curve (blue), reflecting a higher Area Under the Curve (0.931 vs. 0.905). This indicates that the MTL model more reliably distinguishes positive from negative cases across all thresholds.

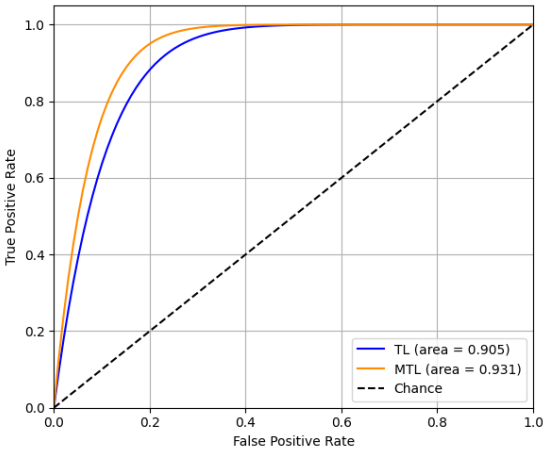


FIGURE 9.2: ROC curves for the MTL and TL models.

9.4.2 False-positive rate and analyst workload

A lower false-positive rate (FPR) reduces the burden on human analysts. The MTL system records an FPR of 4.5 % (BEC) and 2.9 % (AV), compared with 5.5 % and 5.8 % for the TL baseline. All models flag borderline cases to a "red-flag" queue for manual review, preventing critical messages from being silently dropped.

9.4.3 Computational Efficiency

In addition to classification performance, we evaluated the computational efficiency of both the Transfer Learning (TL) and Multi-Task Learning (MTL) models during inference. Table 9.3 reports the average evaluation times on the validation and test sets.

Despite its additional architectural complexity, the MTL model demonstrates slightly faster evaluation times compared to the TL baseline. This improvement is primarily due to the shared encoder being used for both tasks in a single forward pass, whereas the TL setup requires two separate stages—one for BEC detection and a subsequent one for authorship verification. The reduced runtime highlights the practical advantage of deploying a joint model in time-sensitive environments such as real-time email filtering systems.

TABLE 9.3: Computational Efficiency Metrics: Total time in seconds to evaluate the entire validation and test sets.

Model	**Validation Evaluation Time (sec)**	**Test Evaluation Time (sec)**
Transfer Learning	211.85	421.42
Multi-Task Learning	**189.49**	**378.61**

9.5 Analysis of Learned Representations

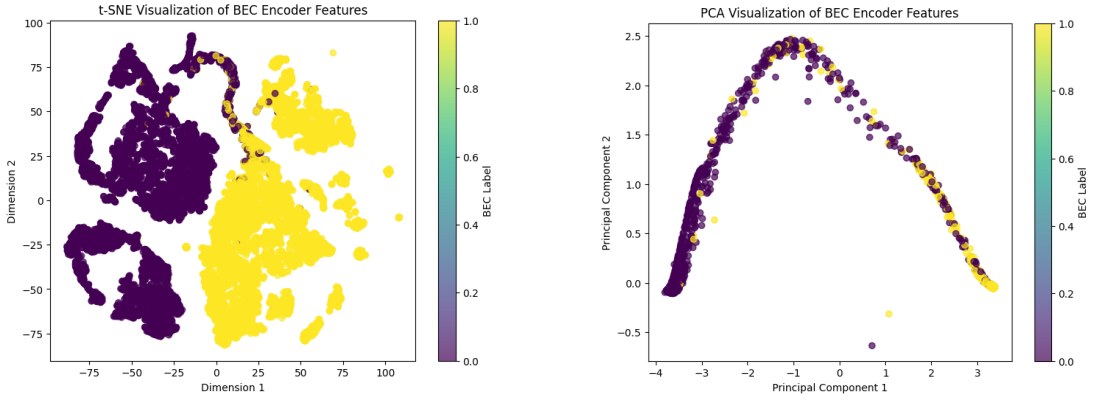
In addition to standard performance metrics, we investigated the *internal* embeddings learned by our MTL model via two popular dimensionality-reduction techniques: **t-SNE** (t-distributed stochastic neighbour embedding) and **PCA** (principal component analysis).

- t-SNE is a nonlinear method that preserves local neighbourhood relationships, often revealing tight clusters that correspond to subtle differences in the data.
- PCA is a linear technique that finds orthogonal axes (principal components) capturing the maximum variance, giving insight into the global structure of the embeddings.

Figures 9.3 and 9.4 show 2D projections of the shared encoder features for the BEC and AV tasks, respectively. In each case, colors encode the true class labels (purple/blue=negative, yellow=positive).

9.5.1 BEC Task Analysis

Figure 9.3a (t-SNE) shows two well-separated clusters of purple (non-BEC) and yellow (BEC) points. This indicates that the shared encoder has learned features—likely things such as vocabulary, phrasing, or tone—that reliably distinguish phishing/impersonation attempts from legitimate emails. In Figure 9.3b (PCA), nearly all purple points lie on one side (negative PC1) and all yellow points on the other (positive PC1), confirming that the first principal component alone captures the majority of the variance correlated with the BEC label. The slight "arc" shape arises because PC2 accounts for a small amount of additional variation, but overall PC1 is highly discriminative.



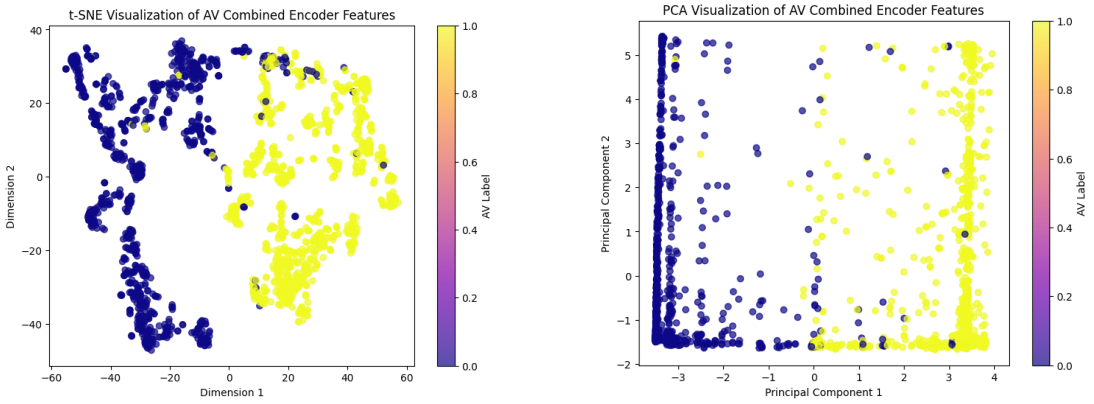
(A) t-SNE of BEC encoder outputs

(B) PCA of BEC encoder outputs

FIGURE 9.3: Dimensionality-reduced embeddings for the BEC task. Each point represents one email's final encoder output. Labels: purple=non-BEC, yellow=BEC.

9.5.2 AV Task Analysis

In the t-SNE plot (Figure 9.4a), points labeled "different author" (blue) and "same author" (yellow) form distinct, well-separated clusters, showing that the encoder captures text-intrinsic cues—such as writing style or vocabulary usage—sufficient to distinguish author pairs. In the PCA projection (Figure 9.4b), almost all blue points lie on the far left (negative PC1) and yellow points on the far right (positive PC1), indicating that PC1 alone already explains a large portion of the variance correlated with authorship similarity.



(A) t-SNE of AV encoder outputs

(B) PCA of AV encoder outputs

FIGURE 9.4: Dimensionality-reduced embeddings for the AV task. Each point is the joint embedding of an email pair. Labels: blue="different author," yellow="same author."

These visualizations were generated from the embeddings of the held-out test set used in Table 4. The fact that BEC vs. non-BEC messages and "same author" vs. "different author" pairs appear as clear clusters under both t-SNE and PCA reinforces our observation that the MTL model's shared encoder learns robust, task-discriminative representations. In other words, the same representations that give rise to the higher accuracy, F1 scores, and lower false-positive rates (compared to the TL baseline) also organize themselves neatly by label when reduced to two dimensions.

9.6 Chapter Summary

This chapter introduced a unified Multi-Task Learning (MTL) framework designed to jointly address Business Email Compromise (BEC) detection and AV. It began by motivating the integration of these two tasks and reviewing related work in multi-task architectures within NLP and cybersecurity. The proposed model combines a shared BERT–BiLSTM encoder with task-specific heads and independent loss functions. We described the training procedure, baseline setup, evaluation metrics, and datasets, including the use of real and synthetic email samples. Experimental results, including classification performance, false-positive rates, and embedding visualizations, were presented to assess the effectiveness of the MTL approach.

Chapter 10

Conclusions and Future Work

10.1 Summary of the Thesis

This thesis presented a unified, NLP-driven framework for detecting Business Email Compromise (BEC) attacks and verifying authorship in enterprise emails. Unlike traditional security tools that rely on metadata or user behavior, the proposed solution focuses exclusively on the content of the email body—using semantic and stylistic cues to identify deception and impersonation. The framework combines three core contributions:

- A transformer–BiLSTM classifier optimized for content-only BEC detection.
- A Siamese authorship verification model (BiBERT-AV) robust against mimicry.
- A joint multi-task learning (MTL) architecture that improves performance and efficiency by learning shared representations across both tasks.

Together, these models demonstrate that semantic deception detection and stylistic verification are complementary components of modern email security—especially when metadata is unavailable or compromised.

10.2 Key Findings

1. **Semantic Models Alone Are Not Enough.** Transformer-based classifiers (e.g., DistilBERT–BiLSTM) perform well on benchmark phishing datasets. However, their accuracy drops significantly under impersonation attacks, where emails are crafted to mimic internal communication styles. This reveals a critical limitation: phishing corpora fail to capture the complexity of BEC threats.

2. **Stylistic Verification Resists Impersonation.** BiBERT-AV—a Siamese network trained to compare writing styles—achieves over 90% accuracy even when faced with dozens of potential authors and style-mimicked messages. It remains robust against adversarial paraphrasing generated using LLMs, making it a valuable defense layer when account takeover occurs.

3. **Multi-Task Learning Boosts Accuracy and Efficiency.** The MTL framework, which shares a common encoder between BEC detection and AV tasks, outperforms both single-task and transfer learning baselines. It also reduces inference time, offering a deployable solution for real-time email filtering that scales with enterprise needs.

10.3 Broader Implications

This research carries important implications for both practice and academic inquiry:

10.3.1 Content-Based Email Security

Defenders should not depend solely on headers, IP addresses, or behavioral signals. When accounts are compromised, the email content remains the only trustworthy signal. NLP-based models like those in this thesis offer a resilient, deployable fallback.

10.3.2 Dual-Gate Filtering

Integrating AV as a secondary check can prevent false negatives by validating whether the writing style matches the claimed sender, especially useful for internal emails or high-risk roles (e.g., executives or finance teams).

10.3.3 Efficient Deployment with MTL

A joint model not only improves accuracy but reduces alert fatigue and latency, supporting proactive rather than reactive defense.

10.3.4 Rethinking Benchmarks

The thesis reinforces that phishing benchmarks are insufficient for BEC evaluation. Realistic assessments must include mimicry, impersonation, and AV-style challenges.

10.4 Limitations

While the findings are promising, several constraints should be acknowledged:

10.4.1 Dataset Limitations

- **Lack of Public BEC Corpora:** Real BEC emails are scarce due to privacy and legal issues.
- **Synthetic Data Caveats:** The thesis uses LLaMA-generated BEC and mimic emails (Chapter 6). While these are validated by human and BLEU/ROUGE scores, they cannot fully replicate adversarial creativity or nuance.

10.4.2 Generalisability

- **Language:** Experiments are limited to English. Results may not generalize to multilingual or code-switched communication.
- **Domain:** Data is based on Enron-style business email. Governmental, legal, or medical contexts may require domain-specific retraining.

10.4.3 Deployment Assumptions

AV assumes access to historical emails per author to compute reference embeddings. In scenarios with new users or limited history, performance may degrade.

10.4.4 Baseline Scope

Only hard-parameter sharing MTL was explored. Variants like soft sharing or hierarchical chaining could yield deeper insights into task synergy.

10.5 Future Work

The thesis opens several avenues for continuation:

1. **Multilingual and Domain-Specific Models**
    - Curate multilingual BEC datasets (e.g., Arabic).
    - Fine-tune models for specialized domains such as finance, legal, or healthcare.

2. **Explainability and Analyst Trust**
    - Apply SHAP/LIME to highlight important tokens.
    - Use counterfactuals to demonstrate how small changes alter predictions—improving transparency and adoption.
3. **Alternative MTL Architectures**
    - Explore soft parameter sharing with task-specific encoders and shared constraints.
    - Investigate task chaining, where AV outputs inform BEC detection.
4. **Multi-Modal and Psycholinguistic Extensions**
    - Integrate non-textual cues: device fingerprinting, metadata, behavioral graphs.
    - Model psycholinguistic traits to strengthen author profiles and detect subtle mimicry.

10.6 Summary of Contributions

TABLE 10.1: Summary of Thesis Contributions

Contribution	**Description**
Systematic Review	Provided the first structured analysis of BEC detection strategies, including non-technical defenses.
BEC Taxonomy	Introduced a five-axis framework grounded in case studies.
BEC Detector	Built and evaluated a transformer–BiLSTM classifier outperforming baselines on phishing and fraud datasets.
AV Model	Designed BiBERT-AV, a mimic-resistant Siamese network using content-only input, scaling to many authors.
Synthetic Dataset	Generated and validated a novel mimicry-aware BEC corpus using LLaMA fine-tuning and human scoring.
Joint MTL Framework	Proposed and validated a multi-task architecture combining semantic and stylistic deception detection with improved speed and accuracy.

10.7 Final Reflections

This thesis has addressed the critical challenge of enhancing email security against sophisticated BEC attacks through advanced NLP techniques. By developing a unified framework that integrates BEC detection and Authorship Verification, it bridges gaps in existing methodologies and provides a foundation for robust, scalable solutions. The findings contribute to both theoretical advancements and practical applications, paving the way for future innovations in combating email fraud.

The hope is that this research will inspire further exploration in the field of email security and encourage the development of intelligent systems capable of adapting to the evolving landscape of cyber threats.

References

Ahmed Abbasi and Hsinchun Chen. Applying authorship analysis to extremist-group web forum messages. *IEEE Intelligent Systems*, 20(5):67–75, 2005.

Abbas Acar, Long Lu, A Selcuk Uluagac, and Engin Kirda. An analysis of malware trends in enterprise networks. In *International Conference on Information Security*, pages 360–380. Springer, 2019.

Rasim Alguliyev, Ramiz Aliguliyev, and Lyudmila Sukhostat. An approach for business email compromise detection using nlp and deep learning. In *2024 IEEE 18th International Conference on Application of Information and Communication Technologies (AICT)*, pages 1–6, 2024. .

Amirah Almutairi, BooJoong Kang, and Nawfal Fadhel. The effectiveness of transformer-based models for bec attack detection. In *International Conference on Network and System Security*, pages 77–90. Springer, 2023.

K Aparna, G Roopesh Kumar, S Ishar, N Santhosh, and D Sreeja. Casestudy on ddos attacks and attack trendsin cloud computing environments. *INTERNATIONAL JOURNAL OF TECHO-ENGINEERING*, 2021.

Emir Araujo-Pino, Helena Gómez-Adorno, and Gibran Fuentes Pineda. Siamese network applied to authorship verification. In *CLEF (Working Notes)*, 2020.

Hany F. Atlam and Olayonu Oluwatimilehin. Business email compromise phishing detection based on machine learning: A systematic literature review. *Electronics*, null:null, 2022. . URL https://www.semanticscholar.org/paper/7d0b27c0830aaaa5932882cb73daa5da528f877d.

Shadrack Awah Buo. An application of cyberpsychology in business email compromise. *arXiv e-prints*, pages arXiv–2011, 2020.

Ronnie T Baby, V Ebenezer, and N Karthik. Magnum opus of phishing techniques. *INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH*, 2019.

M Bagavandas and G Manimannan. Style consistency and authorship attribution: A statistical investigation. *Journal of Quantitative Linguistics*, 15(1):100–110, 2008.

Douglas Bagnall. Author identification using multi-headed recurrent neural networks. *arXiv preprint arXiv:1506.04891*, 2015.

Michel Benaroch. Real options models for proactive uncertainty-reducing mitigations and applications in cybersecurity investment decision making. *Information Systems Research*, 29 (2):315–340, 2018.

Adam Binks. The art of phishing: past, present and future. *Computer Fraud & Security*, 2019(4): 9–11, 2019.

Jan Brabec, Filip Šrajer, Radek Starosta, Tomáš Sixta, Marc Dupont, Miloš Lenoch, Jiří Menšík, Florian Becker, Jakub Boros, Tomáš Pop, et al. A modular and adaptive system for business email compromise detection. *arXiv preprint arXiv:2308.10776*, 2023.

Marcelo Luiz Brocardo, Issa Traore, Sherif Saad, and Isaac Woungang. Authorship verification for short messages using stylometry. In *2013 International Conference on Computer, Information and Telecommunication Systems (CITS)*, pages 1–6. IEEE, 2013.

Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. Signature verification using a" siamese" time delay neural network. *Advances in neural information processing systems*, 6, 1993.

Peter F Brown, John Cocke, Stephen A Della Pietra, Vincent J Della Pietra, Frederick Jelinek, John Lafferty, Robert L Mercer, and Paul S Roossin. A statistical approach to machine translation. *Computational linguistics*, 16(2):79–85, 1990.

AJ Burns, M Eric Johnson, and Deanna D Caputo. Spear phishing in a barrel: Insights from a targeted phishing campaign. *Journal of Organizational Computing and Electronic Commerce*, 29(1):24–39, 2019.

John Burrows. 'delta': a measure of stylistic difference and a guide to likely authorship. *Literary and linguistic computing*, 17(3):267–287, 2002.

Rich Caruana. Multitask learning. *Machine learning*, 28(1):41–75, 1997.

Davide Chicco. Siamese neural networks: An overview. *Artificial neural networks*, pages 73–94, 2021.

Noam Chomsky. Logical structure in language. *Journal of the American Society for Information Science*, 8(4):284, 1957.

Arjun Choudhry, Inder Khatri, Minni Jain, and Dinesh Kumar Vishwakarma. An emotion-aware multitask approach to fake news and rumor detection using transfer learning. *IEEE Transactions on Computational Social Systems*, 11(1):588–599, 2022.

Gobinda G. Chowdhury. Natural language processing. *Annual Review of Information Science and Technology*, 37(1):51–89, 2003. . URL https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/aris.1440370103.

Asaf Cidon, Lior Gavish, Itay Bleier, Nadia Korshun, Marco Schweighauser, and Alexey Tsitkin. High precision detection of business email compromise. In *28th USENIX Security Symposium (USENIX Security 19)*, pages 1291–1307, 2019.

L Cohen. Manion, l.-morrison, k. *Research methods in education*, 5, 2007.

John W Creswell and Vicki L Plano Clark. *Designing and conducting mixed methods research*. Sage publications, 2017.

Cassandra Cross and Rosalie Gillett. Exploiting trust for financial gain: An overview of business email compromise (bec) fraud. *Journal of Financial Crime*, 2020.

Li Deng and Yang Liu. *Deep learning in natural language processing*. Springer, 2018.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Chenhui Dou, Chen Gong, Zhenghua Li, Zhefeng Wang, Baoxing Huai, and Min Zhang. Improving chinese named entity recognition with multi-grained words and part-of-speech tags via joint modeling. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8732–8742, 2024.

Maël Fabien, Esaú Villatoro-Tello, Petr Motlicek, and Shantipriya Parida. Bertaa: Bert fine-tuning for authorship attribution. In *Proceedings of the 17th International Conference on Natural Language Processing (ICON)*, pages 127–137, 2020.

FBI. Operation rewired, Sep 2021. URL https://www.ic3.gov/Media/Y2022/PSA220504.

FBI Internet Crime Complaint Center. 2023 internet crime report. https://www.ic3.gov/Media/PDF/AnnualReport/2023_IC3Report.pdf, 2024. Accessed: 2025-05-21.

FBI Internet Crime Complaint Center (IC3). 2024 internet crime report. https://www.ic3.gov, 2025. Accessed: 2025-01-20.

Federal Bureau of Investigation. Ic3 brochure, March 2024. URL https://www.ic3.gov/Outreach/Brochures/IC3-Brochure.pdf. Accessed: 2025-05-22.

Federal Bureau of Investigation, Internet Crime Complaint Center (IC3). 2023 elder fraud annual report. https://www.ic3.gov/annualreport/reports/2023_ic3elderfraudreport.pdf, 2024. Accessed on May 22, 2025.

Sujata Garera, Niels Provos, Monica Chew, and Aviel D Rubin. A framework for detection and measurement of phishing attacks. In *Proceedings of the 2007 ACM workshop on Recurring malcode*, pages 1–8, 2007.

Hugo Gascon, Steffen Ullrich, Benjamin Stritter, and Konrad Rieck. Reading between the lines: content-agnostic detection of spear-phishing emails. In *International Symposium on Research in Attacks, Intrusions, and Defenses*, pages 69–91. Springer, 2018.

David AE Haddon. Attack vectors and the challenge of preventing data theft. In *CYBER SECURITY PRACTITIONER'S GUIDE*, pages 1–50. World Scientific, 2020.

Oren Halvani, Lars Graner, and Martin Steinebach. On the usefulness of compression models in authorship verification. *Pattern Recognition Letters*, 93:118–126, 2017.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8): 1735–1780, 1997.

Internet Crime Complaint Center (IC3). 2023 internet crime report, 2023. URL https://www.ic3.gov/Media/PDF/AnnualReport/2023_IC3Report.pdf. Accessed: 2024-05-24.

Quanliang Jing, Di Yao, Xinxin Fan, Baoli Wang, Haining Tan, Xiangpeng Bu, and Jingping Bi. Transfake: multi-task transformer for multimodal enhanced fake news detection. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2021.

Vlado Kešelj, Fuchun Peng, Nick Cercone, and Calvin Thomas. N-gram-based author profiles for authorship attribution. In *Proceedings of the conference pacific association for computational linguistics, PACLING*, volume 3, pages 255–264, 2003.

Douglas King. The future of us fraud in a post-emv environment. *Federal Reserve Bank of Atlanta*, 2019.

Barbara Kitchenham and Stuart Charters. Guidelines for performing systematic literature reviews in software engineering. *EBSE*, 2007.

Jan Kolouch. *CyberCrime*. CZ. NIC, zspo, 2016.

Moshe Koppel and Jonathan Schler. Authorship verification as a machine learning task. In *Proceedings of the 14th CPAIOR*, 2004.

Moshe Koppel, Jonathan Schler, and Shlomo Argamon. Authorship attribution in the wild. *Language Resources and Evaluation*, 45:83–94, 2011.

Kroll. Business email compromise attack investigation and remediation for insurance broker, 2021. URL https://www.kroll.com/en/insights/publications/cyber/case-studies/business-email-compromise-attack-investigation. Accessed: 2024-06-05.

Rina Kumari, Nischal Ashok, Tirthankar Ghosal, and Asif Ekbal. Misinformation detection using multitask learning with mutual learning for novelty detection and emotion recognition. *Information Processing & Management*, 58(5):102631, 2021.

Masaki Kurematsu, Ryuhei Yamazaki, Ryo Ogasawara, Jun Hakura, and Hamido Fujita. A study of email author identification using machine learning for business email compromise. In *SoMeT*, pages 205–216, 2019.

Suleman Lazarus. Cybercriminal networks and operational dynamics of business email compromise (bec) scammers: Insights from the "black axe" confraternity. *Deviant Behavior*, pages 1–25, 2024.

Raymond Lister. Mixed methods: positivists are from mars, constructivists are from venus. *ACM SIGCSE Bulletin*, 37(4):18–19, 2005.

Craig Macdonald, Iadh Ounis, and Ian Soboroff. Overview of the trec 2007 blog track. In *TREC*, volume 7, pages 31–43, 2007.

Nasim Maleki. *A behavioral based detection approach for business email compromises*. PhD thesis, University of New Brunswick., 2019.

Steve Mansfield-Devine. The imitation game: How business email compromise scams are robbing organisations. *Computer Fraud & Security*, 2016(11):5–10, 2016.

Thomas Corwin Mendenhall. The characteristic curves of composition. *American Association for the Advancement of Scienc*, 9(214s):237–246, 1887.

Adam Meyers. Not your fairy-tale prince: the nigerian business email compromise threat. *Computer Fraud & Security*, 2018(8):14–16, 2018.

NACHA. Fbi's ic3 finds almost $8.5 billion lost to business email compromise in last three years. https://www.nacha.org/news/fbis-ic3-finds-almost-85-billion-lost-business-email-compromise-last-three-years, 2024. Accessed: 2025-01-20.

Alaa Nehme and Joey F George. Iterating the cybernetic loops in anti-phishing behavior: A theoretical integration. *Twenty-fourth Americas Conference on Information Systems*, 2018.

Robert C Nickerson, Upkar Varshney, and Jan Muntermann. A method for taxonomy development and its application in information systems. *European journal of information systems*, 22(3): 336–359, 2013.

Okechukwu Ogwo-Ude. Business email compromise challenges to medium and large-scale firms in usa: An analysis. *Open Journal of Applied Sciences*, 13(6):803–812, 2023.

Bridget Opazo, Don Whitteker, and Chen-Chi Shing. Email trouble: Secrets of spoofing, the dangers of social engineering, and how we can help. In *2017 13th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)*, pages 2812–2817. IEEE, 2017.

Ivan et al. Ordoñez. Will long-document transformers help authorship verification? In *Working Notes of CLEF 2020*, 2020.

Anastasios Papathanasiou, Georgios Germanos, Nicholas Kolokotronis, and Euripidis Glavas. Cognitive email analysis with automated decision support for business email compromise prevention. In *2023 8th South-East Europe Design Automation, Computer Engineering, Computer Networks and Social Media Conference (SEEDA-CECNSM)*, pages 1–5. IEEE, 2023.

Anastasios Papathanasiou, George Liontos, Georgios Paparis, Vasiliki Liagkou, and Euripides Glavas. Bec defender: Qr code-based methodology for prevention of business email compromise (bec) attacks. *Sensors*, 24(5):1676, 2024.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 2019.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Hao et al. Peng. Encoding word sequences for open-set authorship verification. In *Proceedings of ACL 2021*, 2021.

Flor Miriam Plaza-Del-Arco, M Dolores Molina-González, L Alfonso Ureña-López, and María Teresa Martín-Valdivia. A multi-task learning approach to hate speech detection leveraging sentiment analysis. *IEEE Access*, 9:112478–112489, 2021.

Nathalie Potha and Efstathios Stamatatos. A profile-based method for authorship verification. *Pattern Recognition Letters*, 43:58–67, 2014.

Zheng Qu, Chen Lyu, and Chi-Hung Chi. Multi-task learning framework for detecting hashtag hijack attack in mobile social networks. In *2022 IEEE 19th International Conference on Mobile Ad Hoc and Smart Systems (MASS)*, pages 90–98. IEEE, 2022.

D Radev. Clair collection of fraud email, acl data and code repository. *ADCR2008T001*, 2008.

Mehdi Regina, Maxime Meyer, and Sébastien Goutal. Text data augmentation: Towards better detection of spear-phishing emails. *arXiv preprint arXiv:2007.02033*, 2020.

Chris Ross. The latest attacks and how to stop them. *Computer Fraud & Security*, 2018(11): 11–14, 2018.

Sebastian Ruder. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*, 2017.

Sebastian Ruder, Parsa Ghaffari, and John G Breslin. Character-level and multi-channel convolutional neural networks for large-scale authorship attribution. *arXiv preprint arXiv:1609.06686*, 2016.

Per Runeson and Martin Höst. Guidelines for conducting and reporting case study research in software engineering. *Empirical software engineering*, 14(2):131–164, 2009.

Prasanta Kumar Sahoo and Cheguri Rajitha. Detecting forged e-mail using data mining techniques. *International Journal of Engineering and Advanced Technology*, 2019.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

Norah Saud Al-Musib, Faeiz Mohammad Al-Serhani, Mamoona Humayun, and N.Z. Jhanjhi. Business email compromise (BEC) attacks. *Materials Today: Proceedings*, 3(xxxx), 2021. ISSN 22147853. . URL https://doi.org/10.1016/j.matpr.2021.03.647.

Hanane Sebbaq et al. Mtbert-attention: An explainable bert model based on multi-task learning for cognitive text classification. *Scientific African*, 21:e01799, 2023.

Microsoft Security. What is business email compromise?, 2017. URL https://www.microsoft.com/en-us/security/business/security-101/what-is-business-email-compromise-bec-areaheading-oc3a35.

FBI Public Service-Announcement. Business email compromise: The $55 billion scam. https://www.ic3.gov/Media/Y2024/PSA240911, 2024. Accessed: 2025-05-21.

Vahid Shahrivari, Mohammad Mahdi Darabi, and Mohammad Izadi. Phishing detection using machine learning techniques. *arXiv preprint arXiv:2009.11116*, 2020.

Hossein Siadati. *Prevention, detection, and reaction to cyber impersonation attacks*. PhD thesis, New York University Tandon School of Engineering, 2019.

Hossein Siadati, Jay Koven, Christian Felix da Silva, Markus Jakobsson, Enrico Bertini, David Maimon, and Nasir Memon. A framework for analysis attackers' accounts. In *Security, Privacy and User Interaction*, pages 63–89. Springer, 2020.

Michael Spangler. *Business Email Compromise: Impacts and Strategies for Protecting Against Social Engineering Attacks*. PhD thesis, Utica College, 2021.

Efstathios Stamatatos. A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology*, 60(3):538–556, 2009.

Statista. Number of sent and received e-mails per day worldwide from 2017 to 2026, 2024. URL https://www.statista.com/statistics/456500/daily-number-of-e-mails-worldwide/. Accessed May 2025.

Dwi Siska Susanti, Fitria Errinandini Subandi, Naila Failasufa, and Wibi Anska Putri. Business email compromise (bec) fraud and how to prevent it. *Asia Pacific Fraud Journal*, 8(2):269–280, 2023.

George R Taylor. *Integrating quantitative and qualitative methods in research*. University press of America, 2005.

Songpon Teerakanok, Hiroaki Yasuki, and Tetsutaro Uehara. A Practical Solution against Business Email Compromise (BEC) Attack using Invoice Checksum. *Proceedings - Companion of the 2020 IEEE 20th International Conference on Software Quality, Reliability, and Security, QRS-C 2020*, pages 160–167, 2020. .

Tessian. Business email compromise examples, 2021. URL https://www.tessian.com/blog/business-email-compromise-bec-examples/. Accessed: 2024-06-05.

Wiem Tounsi and Helmi Rais. A survey on technical threat intelligence in the age of sophisticated cyber attacks. *Computers & security*, 72:212–233, 2018.

Jacob Tyo, Bhuwan Dhingra, and Zachary C Lipton. Siamese bert for authorship verification. In *CLEF (Working Notes)*, pages 2169–2177, 2021.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

Alisa Vorobeva, Guldar Khisaeva, Danil Zakoldaev, and Igor Kotenko. Detection of business email compromise attacks with writing style analysis. In *International Symposium on Mobile Internet Security*, pages 248–262. Springer, 2021.

Mo Wang. Profiling retirees in the retirement transition and adjustment process: examining the longitudinal change patterns of retirees' psychological well-being. *Journal of applied psychology*, 92(2):455, 2007.

Wenhui Wang, Nan Yang, Furu Wei, Baobao Chang, and Ming Zhou. Gated self-matching networks for reading comprehension and question answering. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 189–198, 2017.

Ty Wickline. *The Capabilities of Antivirus Software to Detect and Prevent Emerging Cyberthreats*. PhD thesis, Utica College, 2021.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.emnlp-demos.6.

Da Xiao and Meiyi Jiang. Malicious mail filtering and tracing system based on knn and improved lstm algorithm. In *2020 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing,*

Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech), pages 222–229. IEEE, 2020.

Yanling Xiao, Lemao Liu, Guoping Huang, Qu Cui, Shujian Huang, Shuming Shi, and Jiajun Chen. Bitiimt: A bilingual text-infilling method for interactive machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1958–1969, 2022.

Qi Yang and Lin Shang. Multi-task learning with bidirectional language models for text classification. In *2019 international joint conference on neural networks (IJCNN)*, pages 1–8. IEEE, 2019.

Yong Yu, Xiaosheng Si, Changhua Hu, and Jianxun Zhang. A review of recurrent neural networks: Lstm cells and network architectures. *Neural computation*, 31(7):1235–1270, 2019.

Mei Zuo and Yang Zhang. Dataset-aware multi-task learning approaches for biomedical named entity recognition. *Bioinformatics*, 36(15):4331–4338, 2020.

David Zweighaft. Business email compromise and executive impersonation: are financial institutions exposed? *Journal of Investment Compliance*, 2017.