CHALMERS
UNIVERSITY OF TECHNOLOGY

UNIVERSITY OF GOTHENBURG

Explainable AI for Automatic Document Classification in Regulated Finance

Master's thesis in Computer science and engineering

Zachris Stenhammar

Praveen Alavala

Department of Computer Science, Algorithms, Languages and Logic (MPALG)
Department of Computer Science, University of Gothenburg (N2COS)
Chalmers University of Technology and University of Gothenburg
Gothenburg, Sweden 2026

Master's Thesis 2026

Explainable AI for Automatic Document Classification in Regulated Finance

Zachris Stenhammar
Praveen Alavala

GÖTEBORGS UNIVERSITET

CHALMERS

Department of Computer Science – Algorithms, Languages and Logic (MPALG)
Chalmers University of Technology

Department of Computer Science (N2COS)
University of Gothenburg

Gothenburg, Sweden 2026

Explainable AI for Automatic Document Classification in Regulated Finance
Zachris Stenhammar
Praveen Alavala
Department of Computer Science, Algorithms, Languages and Logic (MPALG)
Chalmers University of Technology

Department of Computer Science, University of Gothenburg (N2COS)

Supervisor: Inari Listenmaa, Computer Science and Engineering
Industry Supervisor: Anders Markström, Norion Bank
Examiner: Aarne Ranta, Computer Science and Engineering

Master's Thesis 2026
MPALG and N2COS
Chalmers University of Technology and University of Gothenburg
Gothenburg, Sweden 2026
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Acknowledgements, dedications, and similar personal statements in this thesis, reflect the author's own views.

Typeset in LaTeX
Gothenburg, Sweden 2026

Abstract

The increasing volume of digital documents in regulated financial environments has created significant challenges related to information security, regulatory compliance, and efficient information management. Financial institutions routinely process sensitive information, including internal business data, customer records, and regulatory documents, where incorrect handling or classification may result in legal, financial, and reputational consequences. Despite the importance of information classification, the process is often performed manually, making it inconsistent, time-consuming, and difficult to scale. These challenges motivate the need for automated and trustworthy document classification systems that can support regulated organizations while maintaining transparency and accountability.

This thesis investigates the use of transformer-based language models for automatic document classification in regulated financial environments, with a particular focus on explainability and auditability. The study explores how contextual and semantic information within documents can be used to distinguish between different information sensitivity levels, including Public, Internal, Confidential, and Strictly Confidential classifications. To address privacy and regulatory constraints, the work utilizes a synthetic and semi-controlled dataset .generated using a controlled template-based synthetic document generation methodology with constrained vocabulary, document structures, and contextual patterns designed to reflect the structural and linguistic characteristics of financial documents while avoiding the use of sensitive real-world data.

The proposed system is based on a fine-tuned transformer architecture combined with explainable artificial intelligence (XAI) techniques. Attention-based explanations and Integrated Gradients feature attribution methods are integrated into the classification pipeline to provide insight into the model's decision-making process. The explainability analysis investigates whether the generated explanations align with meaningful contextual indicators associated with document sensitivity and whether they can support transparency, trust, and compliance requirements within regulated financial settings.

The experimental results demonstrate that transformer-based models can effectively learn contextual patterns related to information sensitivity within the controlled dataset while also providing interpretable explanations of classification decisions. The study further analyzes explanation consistency, confidence behaviour, robustness against external documents, and potential shortcut learning effects. Since both the training and evaluation data were generated using the same controlled template-based document generation methodology, the results should be interpreted within the context of this experimental setting. Although separate documents were used for training and evaluation, come from the same dataset share similar linguistic and structural characteristics. Therefore, further evaluation using independent datasets is required to assess the generalizability

of the proposed approach.

This work contributes to the growing field of explainable AI in regulated industries by demonstrating how modern natural language processing techniques can be combined with explainability methods to support secure, transparent, and trustworthy information classification in financial organizations, while also highlighting the importance of independent evaluation when using controlled and synthetic data.

Contents

1 Introduction

The growing use of digital documents in the regulated industry has posed considerable challenges with regard to information security, compliance and efficient information management. For example, the banking and finance industry requires that sensitive information be managed in compliance with strict regulatory guidelines, while at the same time ensuring efficiency. This section provides the context and rationale behind the study, outlines the research objectives and formulates the problem addressed, along with its limitations.

1.1 Context

In regulated industries like finance or banking, organisations deal with a lot of sensitive information. This could be, for instance, business secrets like strategic reports or product documentation, or customer information like financial details or personal data. The processing and storage of such information must comply with strict regulatory frameworks, particularly the General Data Protection Regulation (GDPR) [1] and increasingly the European Union Artificial Intelligence Act (AI Act) [2]. The correct security classification of each document is a must to allow organisations to protect their business and their customers' privacy.

In practice, document classification is usually performed manually by employees. This is a labor-intensive and inconsistent approach that is based on human judgment, which often results in the use of default labels. As a result, documents may be over- or under-classified, which may limit information sharing and efficiency, or create security and privacy issues, respectively. These issues point to the need to find solutions that can help in information classification.

Recent advances in natural language processing (NLP) have shown that it is possible to apply machine learning techniques to effectively process and classify text-based information. In particular, transformer-based architectures, such as BERT[3], have achieved remarkable results in document classification tasks by leveraging contextual and semantic relationships within text-based data. Similar approaches have been applied to large-scale text classification and financial language modelling, including domain-specific adaptations such as FinBERT for financial text analysis [4]. However, these studies have primarily concentrated on improving the overall predictive capabilities of such techniques while paying minimal attention to the unique needs of regulated industries.

In addition, one of the main challenges faced by financial institutions is to demonstrate transparency in AI-based decision-making processes. The provisions regarding automated decision-making in the regulations, such as Article 22 of the GDPR [1] and the provisions related to transparency and human oversight in the AI Act [2] place a burden of justification on organisations that are engaged in automated processing activities. Nevertheless, complex machine

learning models have been accused of being "black boxes" with low levels of interpretability regarding the functioning of the decision-making process [5]. Explainable Artificial Intelligence (XAI) methods have been introduced to address the problems of interpretability in the functioning of such complex machine learning models.

This study is carried out in collaboration with Norion Bank, which provides an industrial context to the study. The combination of automated document classification, explainability and strict compliance constraints defines the context of this work and motivates the exploration of AI-based solutions that are not only accurate, but also transparent, trustworthy and suitable for deployment in regulated financial environments.

1.2 Research Objectives

The goal of this thesis is to develop an AI-based approach to automatic information classification within regulated industries, specifically within the financial and banking sectors. It also aims to bridge the gap between the advancements made in natural language processing and the needs of information security, regulatory compliance and trust within the organisation.

Security level identification One of the key research objectives is to investigate the potential of transformer-based language models to represent different security levels within information and distinguish between them. This also includes the investigation of the potential of contextual information, vocabulary used within the document and the document structure itself to represent different security levels such as information classified as public, internal, confidential, or restricted.

Explainable classification Another key research objective is to investigate the potential of integrating explainability within the information classification process. This thesis also aims to investigate the potential of using XAI techniques to provide meaningful explanations within the information classification process.

From a technical perspective, the objectives of this study include adapting and fine-tuning existing transformer models in a secure manner, specifically for domain-specific document classification, using synthetic data. This objective also involves assessing the impact of explainability techniques on the trustworthiness of the results, in addition to predictive results and the applicability of these techniques without compromising security or privacy.

Industrial applicability Finally, the study aims to develop and evaluate a prototype in collaboration with Norion Bank. The objective is to not only

demonstrate technical viability, but also to assess the applicability of the approach in a real-world, regulated organisational context.

Explainability research question In addition to these objectives, the study is also motivated by the following research question: *How can explainability mechanisms in AI-based document classification systems support transparency, trust and auditability in a regulated financial context?* This is the research question that frames the evaluation of explainability techniques in relation to their applicability

1.3 Problem Statement

Accurate classification of documents according to information sensitivity is a critical requirement in regulated financial environments. Incorrect classification can result in severe consequences, including data breaches, regulatory non-compliance and loss of customer trust. Despite its importance, document classification is commonly performed manually, relying on individual judgement and default security labels. This approach is inherently error-prone, inconsistent and difficult to scale as document volumes increase.

Automated document classification using artificial intelligence has the potential to improve consistency and efficiency by analysing textual content and contextual information. However, existing AI-based classification approaches are often designed primarily for predictive performance and do not sufficiently address the requirements of regulated industries. In particular, many models operate as black boxes, offering limited insight into how classification decisions are made.

In regulated financial contexts, the lack of transparency poses a significant problem. Financial institutions must be able to justify automated decisions to internal stakeholders, auditors and regulatory authorities. Without clear explanations of why a document is assigned a particular sensitivity level, AI-based systems risk being unsuitable for compliance-driven environments, regardless of their accuracy.

The core problem addressed in this study is therefore the tension between automation and accountability: how to leverage advanced machine learning techniques for document classification while ensuring that the resulting system is transparent, explainable and suitable for audit and regulatory review. Addressing this problem requires not only effective classification models, but also the integration of explainability mechanisms that support trust, oversight and compliance in a regulated financial setting.

1.4 Limitations

This section defines the specific scope and boundaries of the study, highlighting the assumptions and constraints under which the proposed AI-based document

classification system is developed and evaluated. While the Introduction and Context sections provide a broad overview of the problem and its relevance, this section focuses on the practical limitations that shape the methodology, data selection and evaluation. Clarifying these limits helps ensure that the analysis is both feasible and meaningful within the time frame and regulatory context of the study.

The study only considers documents written in English and does not include multilingual document classification. Although multilingual environments are common in financial organisations, restricting the analysis to a single language allows for in-depth evaluation of the model and explainability techniques within a controlled scope.

The study focuses on internal reports. Other common document types are excluded from the analysis. Limiting the study to these specific document types ensures that results are less influenced by variations in structure, style and use across diverse document formats, thereby supporting a more focused and detailed evaluation.

The evaluation of explainability techniques is also constrained to attention-based indicators and Integrated Gradients feature attributions [6]. These techniques are selected because they enable meaningful inspection of the model's decision-making process while remaining compatible with the security and regulatory requirements of the financial organisation. The study does not attempt to analyse or compare all available explainability methods, as doing so would exceed the study's scope. By focusing on a small set of complementary methods, the study ensures thorough and consistent evaluation within the defined time frame.

While the use of synthetic data enables controlled experimentation, it also introduces limitations related to linguistic diversity and realism. The dataset does not fully capture the variability, ambiguity and noise present in real-world financial documents. As a result, model performance and explainability outcomes observed in this study may not fully generalize to unconstrained natural language settings.

2 Theoretical Background

The increasing digitization of organisational processes has resulted in a rapid growth in the volume of textual information processed by financial institutions. Banks and other regulated organisations routinely manage large amounts of textual data originating from internal reports, regulatory filings, customer communications, governance documentation and operational processes. As the volume and complexity of such information continue to increase, organisations face growing challenges related to the secure handling, classification and governance of sensitive documents.

Within regulated financial environments, information classification is an important process used to protect sensitive information, support regulatory compliance and reduce the risks associated with unauthorized disclosure. Traditionally, document classification has often relied on manual review processes or rule-based systems. However, manual classification is typically time-consuming, inconsistent and dependent on individual judgement, while rule-based approaches are often difficult to maintain and adapt to evolving document structures and regulatory requirements.

Recent advances in natural language processing (NLP) and machine learning have created new opportunities for automated document classification. In particular, transformer-based language models such as BERT have demonstrated strong capabilities for modelling contextual and semantic relationships within complex documents. At the same time, the increasing complexity of deep learning models has introduced challenges related to transparency, interpretability and accountability, especially in regulated and high-stakes environments where automated decisions must be understandable and auditable.

Explainable Artificial Intelligence (XAI) has emerged as an important research area aimed at improving transparency and human understanding of machine learning systems. XAI approaches are commonly divided into inherently interpretable models and post-hoc explanation methods. Since transformer-based language models are highly complex and not inherently interpretable, this study primarily focuses on post-hoc explainability techniques that enable inspection and analysis of model predictions without modifying the underlying architecture.

This chapter presents the theoretical foundations relevant to the study. It introduces information classification in regulated financial environments, transformer-based language models and self-attention mechanisms, contextual text representations and explainable artificial intelligence techniques relevant to document classification and regulatory compliance.

2.1 Information Classification in Regulated Financial Environments

Within regulated financial environments, documents are commonly categorised according to predefined levels of information sensitivity in order to support the safeguarding of sensitive information, ensure compliance with regulatory requirements and control access to organisational data. Information classification policies define how documents may be stored, processed, transmitted and shared within an organisation while also reflecting the potential consequences associated with unauthorized disclosure.

Financial institutions process large amounts of sensitive and confidential information on a daily basis, including customer information, internal operational material, legal documentation, financial reports and compliance-related communication. Failures in information classification may lead to information leakage, regulatory violations, reputational damage and financial consequences under regulations such as the EU General Data Protection Regulation (GDPR), the Gramm-Leach-Bliley Act (GLBA) and the Markets in Financial Instruments Directive II (MiFID II).

In practice, information sensitivity is often context-dependent and may not be identifiable through isolated keywords alone. The classification of documents therefore requires consideration of semantic meaning, contextual relationships, domain-specific terminology and document structure. These characteristics make document classification particularly suitable for modern natural language processing approaches capable of modelling contextual information across complex textual documents.

The classification scheme used in this study was provided by the collaborating bank and serves as the foundation for the dataset construction, labelling methodology and evaluation process used throughout the thesis.

2.1.1 Information Classification Levels

Within regulated financial environments, documents are commonly categorised according to predefined levels of information sensitivity in order to support the safeguarding of sensitive information, ensure compliance with applicable regulatory requirements and control access to sensitive information. Classification levels define how an institution may store, process and share documented information, while also indicating the potential consequences associated with unauthorized disclosure.

The classification scheme used in this study was provided by the collaborating bank and served as the foundation for the document labelling methodology used throughout the dataset and evaluation process.

Table 1 provides an overview of the classification scheme used in this study, including descriptions of the information associated with each classification category together with representative examples of information corresponding to each classification level.

Table 1: Information classification levels used in the dataset

Classification	Description	Example Information
Public	Information intended for external parties or the general public. Disclosure of this information has no negative consequences for the Bank.	Publicly available information, press material, recruitment and career information
Internal	Information primarily intended for internal use within the Bank but which may occasionally be shared externally if required. Unauthorized disclosure may lead to minor consequences.	Internal communication, governance documents, work-related operational information
Confidential	Information that should only be shared with a limited list of recipients. Unauthorized disclosure may lead to negative business consequences or regulatory issues.	Salary information, supplier information, business contracts, personal data, production code
Strictly Confidential	Information that should only be shared with a strictly limited list of recipients. Unauthorized disclosure may result in severe consequences such as financial loss, reputational damage, or breaches of laws and regulations.	Insider information, unpublished financial results, banking secrecy information, sensitive personal data

2.1.2 Regulatory and Compliance Requirements

In regulated financial environments, information classification is closely connected to legal, regulatory and organisational compliance requirements. Financial institutions process large amounts of sensitive and confidential information on a daily basis, including customer data, internal business information, financial reports and regulatory documentation. The handling of such information is governed by strict regulations intended to protect privacy, ensure accountability and reduce the risks associated with unauthorized disclosure or misuse of sensitive data.

Failures in information classification may result in information leakage and can lead to severe legal, financial and reputational consequences under regulations such as the EU General Data Protection Regulation (GDPR) [1], the US Gramm-Leach-Bliley Act (GLBA) [7] and the Markets in Financial Instruments Directive II (MiFID II) [8]. These regulations impose obligations related to data protection, confidentiality, transparency and governance of information processing activities.

The GDPR places particular emphasis on the protection of personal data and introduces requirements related to accountability, traceability and automated decision-making. Article 22 of the GDPR specifically addresses automated decision-making processes and highlights the importance of human oversight and the ability to justify automated decisions [1]. In practice, this creates increasing pressure on organisations to ensure that AI-based systems used for document processing and classification remain transparent and explainable.

Similarly, the Gramm-Leach-Bliley Act (GLBA) requires financial institutions to protect customer financial information and implement safeguards for sensitive data [7]. Information classification therefore becomes an essential mechanism for determining how information should be stored, processed and shared within the organisation. In addition, MiFID II introduces extensive requirements regarding governance, record keeping, transparency and auditability within financial operations [1]. These requirements further increase the importance of maintaining clear and well-documented information management processes.

organisations operating in regulated domains are frequently required to provide transparency and accountability for automated decision-making systems in order to satisfy regulatory and governance requirements [9]. Consequently, they must be able to justify decisions made by automated systems to internal stakeholders, auditors and regulatory agencies while demonstrating compliance with applicable regulations and organisational policies [10]. This requirement is particularly important for AI-based systems that process sensitive information or support compliance-related decision-making.

The increasing adoption of machine learning in regulated environments has introduced a widely discussed trade-off between the predictive performance of complex models and the interpretability of their decision processes [11]. Transformer-based language models and other deep learning approaches often achieve strong predictive performance, but their decision-making processes are frequently regarded as opaque or "black-box" systems [12]. In regulated financial contexts, this lack of transparency creates challenges related to accountability, trust and auditability, as organisations may find it difficult to verify whether automated decisions comply with regulatory requirements or internal policies.

For this reason, explainability is increasingly regarded as a core design require-

ment in trustworthy AI systems rather than merely an optional post-hoc capability [13]. Explainable Artificial Intelligence (XAI) methods enable organisations to inspect, validate and justify the outputs produced by automated systems [12]. In the context of AI-based document classification, explainability mechanisms can help demonstrate that classification decisions are based on meaningful contextual and semantic information rather than irrelevant or potentially biased patterns.

The need for transparency and explainability is therefore not only a technical consideration, but also a regulatory and organisational requirement in modern financial environments. Automated document classification systems intended for deployment in regulated organisations must consequently balance predictive performance with explainability, auditability and human oversight in order to support trustworthy and compliant information management processes.

2.2 Transformer-Based Language Models

Natural language processing has mainly focused on sequential and convolutional neural networks, including recurrent neural networks [14–16]. Although these models have shown promising results in solving various problems in natural language processing tasks, including text classification problems, they have shown some disadvantages when applied to complex documents [17]. One disadvantage of sequential neural networks is that they process text token by token. This makes it difficult to capture contextual relationships within a document [14, 15].

To address these limitations, previous studies proposed hierarchical neural network approaches that can represent a document using a Hierarchical Neural Network [18]. Although this approach can represent a document more effectively compared to sequential neural networks, it still shows some disadvantages associated with sequential neural networks. To overcome these disadvantages, a new language model was proposed that can represent a document by eliminating the need for using recurrence and convolution operations and focusing on using attention mechanisms that allow modelling relationships between all parts of a document [19]. These models have achieved state-of-the-art performance across many text classification tasks [17].

2.2.1 Evolution of Neural NLP Models

Early approaches in natural language processing primarily relied on statistical and rule-based methods that represented text using handcrafted linguistic features and shallow machine learning techniques. Although such approaches proved effective for narrow classification tasks, they often struggled to capture semantic meaning, contextual dependencies and variations in natural language usage, particularly in complex document-level classification problems.

With the advancement of deep learning, neural network architectures became

increasingly prominent in NLP research. Sequential neural network models, particularly recurrent neural networks (RNNs) and Long Short-Term Memory (LSTM) networks, enabled models to process textual information as ordered sequences while retaining information from previous tokens [14–16]. These approaches improved the ability to model contextual information compared with earlier bag-of-words representations and achieved strong performance across several NLP tasks, including sentiment analysis and text classification.

Convolutional neural networks (CNNs) were also applied to NLP tasks by learning local textual patterns through convolution operations over word embeddings [15]. CNN-based approaches demonstrated effectiveness in sentence classification and feature extraction tasks due to their computational efficiency and ability to identify local semantic patterns. However, both recurrent and convolutional architectures showed limitations when applied to long and complex documents, particularly in scenarios where contextual relationships extend across multiple sections of a document [17].

One important limitation of sequential architectures is that they process text token by token, making long-range dependency modelling computationally difficult and less effective [14, 15]. As document length increases, information from earlier parts of the sequence may become diluted or lost, limiting the model's ability to capture broader contextual relationships. This issue is particularly problematic in document-level classification tasks involving legal, financial, or regulatory documents, where important contextual cues may be distributed throughout the document rather than concentrated locally within a sentence.

To address these limitations, previous research proposed hierarchical neural network approaches that represent documents at multiple structural levels, such as words, sentences and paragraphs [18]. Hierarchical Attention Networks and related architectures improved document representation by modelling both local and document-level contextual information. Although these approaches demonstrated improvements compared with purely sequential architectures, they still inherited several limitations associated with recurrent processing and struggled to efficiently model very long contextual dependencies [18].

The introduction of transformer architectures represented a major shift in NLP research by eliminating the need for recurrence and convolution operations and instead relying primarily on attention mechanisms to model relationships between all parts of a document [19]. Unlike sequential models, transformers process tokens in parallel and establish contextual relationships regardless of token distance, enabling more effective modelling of long-range semantic dependencies. This architectural change significantly improved scalability and contextual understanding while enabling efficient training on large-scale text corpora.

Transformer-based language models such as BERT further advanced this paradigm through large-scale pretraining and contextual bidirectional encoding [3]. Rather

than learning task-specific representations from scratch, pretrained transformer models learn general linguistic and semantic representations that can later be adapted to downstream tasks through supervised fine-tuning. This transfer learning approach has achieved state-of-the-art performance across a wide range of NLP applications, including document classification, question answering, legal text analysis and financial language processing [17].

The evolution from sequential neural networks to transformer-based architectures has therefore significantly improved the ability of NLP systems to capture semantic, contextual and structural information within complex documents. These developments form the foundation for modern AI-based document classification systems and provide the theoretical basis for the transformer-based approach used in this study.

2.2.2 Transformer Architecture and Self-Attention

Transformer-based language models were introduced to address several limitations associated with sequential neural network architectures in natural language processing. Unlike recurrent neural networks and convolutional architectures, transformer models eliminate the need for recurrence and convolution operations by relying primarily on attention mechanisms to model relationships between tokens within a document [19]. This architectural design enables the model to process all tokens in parallel while capturing contextual relationships across the entire input sequence.

The transformer architecture consists of stacked encoder and decoder layers built around self-attention mechanisms and feed-forward neural networks [19]. In natural language processing tasks such as document classification, transformer encoders are commonly used to generate contextualized representations of textual input. Through pretraining on large-scale corpora, transformer-based models learn general linguistic and semantic representations that can later be adapted to downstream tasks through supervised fine-tuning [3]. This transfer learning approach has proven highly effective for document classification tasks, including applications involving legal, regulatory and financial documents [17].

One of the most important parts of the architecture of the transformer is the self-attention mechanism. Unlike in sequential models, the self-attention mechanism in the transformer architecture enables the model to process all the tokens in the document simultaneously and to establish relationships among them irrespective of their distance from one another, thereby eliminating the need for recurrence and convolution operations and improving the modelling of long-range dependencies [19].

The self-attention mechanism enables the model to assign weights to the tokens in the text in order to determine which parts of the input are most relevant when generating internal contextual representations [19]. In practice, this

means that the representation of a given token is influenced not only by the token itself, but also by its relationship to surrounding tokens and other relevant parts of the document. This capability is particularly important for complex document-level classification tasks, where semantically important information may be distributed across multiple sections of a document rather than appearing locally within a sentence [17].

Technically, the self-attention mechanism operates by projecting each token representation into three separate vectors commonly referred to as queries, keys and values [19]. Attention scores are computed by measuring the similarity between query and key representations, after which normalized attention weights are applied to the value representations in order to generate contextualized token embeddings. Through this process, the model learns which tokens should receive greater attention when constructing semantic representations of the document.

In addition to self-attention, transformer architectures commonly employ multi-head attention mechanisms, where multiple attention operations are learned in parallel [19]. This allows the model to capture different types of semantic and contextual relationships simultaneously, including syntactic structure, domain-specific terminology and long-range dependencies between different parts of the document. Such capabilities are especially relevant in financial and regulatory environments, where the interpretation of information often depends on contextual relationships distributed across the document.

Transformer-based architectures such as BERT further extend this capability by using bidirectional contextual encoding [3]. Unlike earlier sequential models that process text primarily in a left-to-right or right-to-left manner, bidirectional transformers incorporate contextual information from both directions simultaneously when generating token representations. This enables the model to capture more complete semantic relationships within the text and improves its ability to distinguish between subtle contextual differences.

In regulated financial environments, the sensitivity of financial documents may depend on contextual cues, terminology and relationships established between different parts of the text. Sensitive information may not always be identifiable through isolated keywords alone, but rather through the broader semantic context in which information appears. Transformer-based models are therefore particularly suitable for financial document classification tasks, as their self-attention mechanisms enable effective modelling of contextual and semantic dependencies throughout the document [10, 19].

The ability of transformer architectures to generate contextualized document representations while simultaneously modelling long-range relationships has contributed significantly to their strong performance across modern document classification tasks [17]. Consequently, transformer-based language models provide

an appropriate theoretical and methodological foundation for this study, which investigates explainable AI-based classification of financial documents in regulated environments.

2.2.3 Contextual Text Representations

Transformer-based language models such as BERT generate contextualized word representations, meaning that the representation of a word depends on its surrounding context [3]. These contextual embeddings differ from static word embeddings, as they adapt dynamically to usage within a sentence [3, 19].

This characteristic of the model is particularly useful for applications such as finance and regulations, where the meaning of words may vary according to the domain and the sensitivity of the documents may depend on the usage of words within a particular context. The ability of the model to understand the context of the words enables the model to differentiate between the security levels based on the vocabulary, structure and overall context of the documents [10, 18].

Based on the characteristics, the model has been effective for the classification of documents within the legal and regulatory environments, such as the classification of the EU legislation [17]. Therefore, transformer-based models provide an appropriate theoretical basis for this study, which involves the classification of financial documents within the regulated financial environments.

2.2.4 Transformers in Financial Document Classification

Transformer-based language models have become increasingly important in document classification tasks involving legal, financial and regulatory text due to their ability to model contextual and semantic relationships across long and complex documents [17]. In regulated financial environments, documents often contain highly domain-specific terminology, contextual dependencies and structurally distributed information that cannot easily be captured through traditional keyword-based or shallow machine learning approaches.

Traditionally, document classification within organisations has relied on manual review processes or rule-based systems. Manual classification is often inconsistent, time-consuming and dependent on individual judgement, while rule-based systems are difficult to maintain and adapt to evolving document structures and regulatory requirements. These limitations have motivated the increasing adoption of machine learning techniques for automatic document classification in regulated domains.

Early machine learning approaches relied primarily on handcrafted features combined with shallow classifiers. Later developments introduced deep learning architectures such as recurrent neural networks, convolutional neural networks and hierarchical neural models for document representation [18]. Although these ap-

proaches improved classification performance, they often struggled to efficiently capture long-range contextual dependencies within lengthy financial and regulatory documents [17].

Transformer-based architectures address these limitations through the use of self-attention mechanisms and contextual embeddings [19]. By modelling relationships between all tokens within a document simultaneously, transformer models are able to capture semantic relationships distributed across multiple sections of text. This capability is especially important in financial document classification, where the sensitivity or meaning of information may depend on contextual cues rather than isolated keywords alone.

Models such as BERT generate contextualized token representations that dynamically adapt according to surrounding text [3]. This allows the model to distinguish between different meanings of the same terminology depending on the broader semantic context in which the terms appear. In regulated financial environments, such contextual understanding is particularly important because information sensitivity classifications often depend on subtle linguistic distinctions, domain-specific phrasing and relationships established across different parts of a document.

Pretrained transformer models have also proven particularly effective in domains where access to large labelled datasets is limited. Through large-scale pretraining on public corpora, transformer architectures learn general linguistic and semantic representations that can later be adapted to specialized downstream tasks through supervised fine-tuning [3]. This transfer learning capability is highly relevant in financial and regulatory contexts, where confidentiality constraints often limit the availability of real-world training data.

Previous research has demonstrated the effectiveness of transformer-based models for legal and regulatory document classification tasks, including classification of EU legislation and other compliance-related text processing applications [17]. Domain-specific adaptations such as FinBERT have further demonstrated how pretrained transformer architectures can be adapted to financial language processing tasks involving financial sentiment analysis, risk analysis and domain-specific text understanding [4].

In addition to predictive performance, transformer-based architectures provide opportunities for explainability analysis through attention mechanisms and feature attribution methods. Since transformer models internally assign varying levels of attention to different parts of the input text, they provide useful indicators regarding which contextual relationships influence classification decisions [19]. This is particularly important in regulated financial environments, where organisations may be required to justify automated decisions to auditors, regulators and internal stakeholders.

The use of transformer-based models in regulated financial document classification therefore reflects both technical and organisational requirements. Technically, transformers provide strong contextual modelling capabilities suitable for complex document classification tasks. organisationally, their compatibility with explainability techniques makes them relevant for environments where transparency, accountability and regulatory compliance are essential requirements.

Consequently, transformer-based language models provide an appropriate methodological and theoretical foundation for this study, which investigates explainable AI-based classification of financial documents within regulated financial environments.

2.3 Explainable Artificial Intelligence

Explainable Artificial Intelligence (XAI) refers to the various methods and techniques developed to provide transparency and comprehensibility to the decision-making processes of artificial intelligence systems [12]. As machine learning models have increased in complexity, their decision-making processes have often become opaque. This lack of transparency raises concerns related to trust, accountability and the ability to justify automated decisions [12, 20].

In the field of natural language processing, recent breakthroughs have been noted with the application of deep learning models, particularly the use of the transformer architecture. This has improved the predictive capabilities of the models on a wide range of tasks [3]. However, the decision-making process has been noted as a black box and this is posing challenges, especially in critical domains where automated decisions need to be traceable and auditable. Explainable artificial intelligence has thus been seen as an option to address the challenges posed by the process of decision-making [12, 21].

XAI approaches are commonly categorised as either inherently interpretable models or post-hoc explanation methods [12]. Inherently interpretable models are designed so that their decision-making processes can be directly understood by human users, often through simpler model structures or transparent decision rules [12]. In contrast, post-hoc explanation methods aim to explain the behaviour of complex machine learning models after training without modifying the underlying architecture [12, 22]. Since transformer-based language models achieve high predictive performance at the cost of increased complexity [3], post-hoc explanation techniques are particularly important for enabling transparency and interpretability in modern NLP systems [12].

2.3.1 Interpretability in NLP

The concept of model interpretability in the field of NLP deals with the understanding of how the model processes the input text and the contribution

of different components of the input text towards the prediction made by the model [12]. Traditional NLP models, such as linear classifiers, generally offer higher interpretability due to their lower complexity. In contrast, modern deep learning models like hierarchical neural networks or transformer-based models, the high-dimensional input data cannot be considered an easily interpretable model [3, 18].

The main problem in the interpretability of the transformer-based model arises from the distributed contextual representations and the attention mechanisms used in these models [3]. These models have a high capability for performance by capturing the semantic and contextual information in the input text. But it is not possible to trace the decision made by the model from input data to output data. So it becomes a complex task to identify which components of the input text are affecting the classification decision made by the model, particularly in document-level classification tasks [12].

The concept of interpretability in the field of NLP is a prerequisite for the application of AI-based systems in real-world scenarios, especially in high-stakes or regulated environments where model behaviour must be understood, validated and justified [12].

In regulated and high-stakes environments, interpretability is not merely desirable but is often necessary for the practical deployment of AI-based systems [12, 13]. organisations must be able to validate, justify and audit automated decisions in order to ensure compliance with regulatory requirements, organisational policies and principles of trustworthy AI [1, 2, 9]. Consequently, interpretability plays an important role in supporting trust, accountability and human oversight in NLP-based document classification systems [10, 12].

2.3.2 Post-hoc Explanation Methods

Post-hoc explanation methods are designed to explain model predictions after training, without modifying the underlying model architecture [12]. In natural language processing, these methods are commonly used to identify the contribution of individual tokens, spans of text, or contextual features to a given prediction [23]. Unlike inherently interpretable models, post-hoc approaches enable the analysis of complex neural architectures while preserving their predictive capabilities, making them particularly suitable for transformer-based language models used in document classification tasks [22].

As transformer-based models have become increasingly prevalent in high-stakes and regulated domains, their use in financial document classification, compliance monitoring and sensitive information handling has expanded. However, the growing complexity of these models has introduced significant challenges related to transparency and accountability. Transformer-based architectures are often regarded as black-box systems because their internal decision-making processes

are not directly interpretable by human users [12]. This lack of transparency poses particular concerns in regulated financial environments, where organisations must be able to justify automated decisions to auditors, regulators and internal stakeholders.

In the context of NLP, post-hoc explanation methods are primarily designed to improve understanding of how specific input components contribute to model predictions. These approaches aim to identify which textual elements most strongly influence the classification decision, thereby enabling inspection of the model's behaviour without compromising predictive performance [12, 22]. Such methods are especially important in document-level classification tasks, where sensitivity classifications may depend on contextual relationships distributed across multiple sections of a document rather than on isolated keywords alone.

One important category of post-hoc explanation methods is feature attribution. Feature attribution methods attempt to estimate the contribution of individual input features to a model prediction by measuring how changes in the input affect the model output [12, 24]. In natural language processing, these methods are commonly applied at the token level in order to identify which words or phrases contribute most strongly to a predicted class label. This provides a more fine-grained view of model behaviour and enables direct inspection of the textual elements influencing a classification decision.

In this study, Integrated Gradients is used as the primary feature attribution method [6]. Integrated Gradients is a gradient-based attribution technique designed to measure the contribution of each input feature relative to a baseline input by accumulating gradients along a continuous interpolation path between the baseline and the original input [6]. Compared with standard gradient-based methods, Integrated Gradients has been shown to provide more stable and theoretically grounded attribution results while remaining compatible with complex neural architectures such as transformers [6, 23].

In addition to feature attribution methods, attention-based indicators are also commonly used as post-hoc explanation mechanisms in transformer-based NLP systems. The self-attention mechanism enables the model to assign different weights to tokens when constructing contextual representations, thereby indicating which parts of the input receive greater focus during inference [19]. Attention visualisations can therefore provide insight into how the model distributes contextual focus across a document. However, the validity of attention weights as faithful explanations remains debated within the literature. Jain and Wallace [25] argue that attention weights are not necessarily reliable explanations of model predictions, whereas Wiegreffe and Pinter [26] suggest that attention can still provide meaningful interpretive value under certain conditions.

For this reason, attention-based explanations in this study are treated primarily as qualitative transparency indicators rather than direct causal explanations

of model behaviour. Attention-based indicators are interpreted together with feature attribution methods in order to provide complementary perspectives on the model's decision-making process. Combining multiple explanation techniques may help reduce the limitations associated with individual methods while improving the reliability and robustness of explanation analysis [24, 25].

Post-hoc explanation methods therefore serve as an important bridge between the strong predictive capabilities of transformer-based language models and the practical requirements of transparency, interpretability and auditability in regulated environments. In document classification systems used within finance and banking, these methods enable inspection and validation of model predictions while supporting trust, human oversight and regulatory compliance [12, 20].

2.3.3 Explainability in Regulated Domains

In regulated industries such as finance and banking, explainability is not only desirable but is often a necessity [12]. Financial institutions process large amounts of sensitive and confidential information on a daily basis, making information classification a critical process in such environments. Failures in information classification may result in information leakage and can lead to severe legal, financial and reputational consequences under regulations such as the EU General Data Protection Regulation (GDPR) [1], the US Gramm-Leach-Bliley Act (GLBA) [7] and the Markets in Financial Instruments Directive II (MiFID II) [8].

organisations operating in regulated domains are frequently required to provide transparency and accountability for automated decision-making systems in order to satisfy regulatory and governance requirements [9]. Consequently, they must be able to justify decisions made by automated systems to internal stakeholders, auditors and regulatory agencies while demonstrating compliance with applicable regulations and organisational policies [10]. This requirement is particularly important for AI-based systems that process sensitive information or support compliance-related decision-making.

The increasing adoption of machine learning in regulated environments has introduced a widely discussed trade-off between the predictive performance of complex models and the interpretability of their decision processes [11]. The use of black-box models in high-stakes domains raises challenges related to accountability, trust and auditability [12], as it can be difficult to verify whether automated decisions comply with regulatory requirements or internal policies. For this reason, explainability is increasingly regarded as a core design requirement in trustworthy AI systems rather than merely an optional post-hoc capability [13].

Explainable artificial intelligence (XAI) enables organisations to inspect, validate and justify the outputs produced by automated systems [12]. This is espe-

cially relevant for NLP-based document classification systems used in controlled industries such as finance and banking, where explainability can help validate sensitivity assessments and ensure that classifications are based on appropriate contextual and semantic factors [10]. Recent research has also explored hybrid NLP and machine learning approaches for identifying and anonymizing sensitive information in financial documents, further emphasising the importance of transparency and regulatory compliance in financial AI systems [21].

3 Related Works

This chapter reviews prior research relevant to explainable AI-based document classification in regulated financial environments. Building upon the theoretical foundations presented in Chapter 2, the chapter focuses on how previous studies have approached financial document classification, explainability in NLP systems and the use of controlled or synthetic datasets for sensitive-domain machine learning.

3.1 Document Classification and XAI in Fintech

This thesis builds upon the transformer-based and explainable NLP approaches discussed in Section 2 by investigating how explainability techniques can be integrated into document classification systems for regulated financial environments. In particular, the study focuses on combining transformer-based classification with post-hoc explainability methods in a controlled financial document classification setting.

Prior research has demonstrated the effectiveness of transformer-based language models for document classification tasks involving legal, regulatory and financial text. Studies involving BERT-based architectures and domain-specific adaptations such as FinBERT have shown strong performance in financial language processing tasks including sentiment analysis, compliance monitoring and regulatory text classification.

In parallel, explainability techniques have increasingly been applied to NLP systems operating in high-stakes domains. Attention visualisation, gradient-based attribution methods and post-hoc explanation techniques have been explored as mechanisms for improving transparency and interpretability in transformer-based systems.

However, much of the existing research primarily focuses on predictive performance rather than explainability, auditability, or regulatory applicability. Furthermore, relatively limited work has investigated explainable document classification specifically for information sensitivity classification within regulated financial environments. Existing studies also frequently rely on either public benchmark datasets or unrestricted natural language corpora, which may not adequately reflect the confidentiality constraints present in real-world financial organisations.

This thesis addresses these gaps by combining transformer-based document classification with post-hoc explainability techniques in a controlled financial document classification setting designed to reflect regulatory and organisational constraints.

3.2 Controlled Natural Languages and Synthetic Data

Prior research on Controlled Natural Languages (CNLs) has explored how constrained linguistic structures can improve consistency, interpretability and machine processing of textual data [27]. Such approaches have commonly been applied in domains requiring precision and reduced ambiguity, including technical documentation and regulatory communication.

Previous studies have shown that reducing linguistic variability may improve the reliability and interpretability of NLP systems operating in controlled environments [12, 17]. These observations are particularly relevant for explainability-oriented NLP research, where excessive linguistic variability may complicate interpretation of model behaviour and attribution results.

Although this study does not implement a formal CNL framework, the dataset design follows several principles associated with controlled language approaches. This includes constrained vocabulary, structured document patterns and reduced linguistic variability. This results in a controlled linguistic environment that reduces variability while maintaining the essential characteristics related to the classification of information sensitivity.

This approach should be distinguished from fully developed CNL systems, which often use formal grammars and tools such as the Grammatical Framework (GF) [28] to impose strict syntactic and semantic constraints [27]. In contrast, this study's method uses lightweight template-based generation and does not guarantee formal linguistic correctness or completeness. This trade-off enables the creation of datasets with reduced linguistic variability while remaining flexible and practical to construct.

Future research may investigate the use of grammar-based generation techniques. This could increase linguistic diversity while maintaining control over grammatical structure, enabling more robust evaluation of model performance and explainability under varying syntactic conditions, while still allowing systematic verification and interpretation of model behaviour.

4 Methodology

This chapter presents the methodological design and implementation choices used to develop and evaluate the proposed explainable AI-based document classification system. The chapter describes the dataset construction process, model implementation, explainability integration and evaluation methodology.

4.1 Dataset

Due to regulatory, confidentiality and privacy constraints associated with financial documents, the study uses a synthetic dataset designed to reflect realistic linguistic and structural characteristics relevant to information sensitivity classification.

The dataset consists of documents assigned to predefined information sensitivity classes corresponding to the classification scheme introduced in Section 2.1.1. Classification is conducted according to explicit labelling criteria in order to reduce ambiguity and subjectivity, acknowledging that information sensitivity is inherently context-dependent [12].

The following subsections describe the construction of the dataset in more detail, including the use of synthetic and anonymized data, as well as the preprocessing and labelling procedures applied prior to model training.

The dataset is constructed as a controlled linguistic environment with constrained variability in vocabulary, structure and contextual patterns in order to support explainability analysis and reduce noise during classification.

4.1.1 Synthetic Data

Synthetic documents are used to represent internal documents. These documents are generated using a template-based approach, where document structure, stylistic conventions and domain-specific terminology are modelled and populated with randomised placeholders and context-appropriate terms. This method enables the simulation of realistic document flow without including any identifiable or confidential information.

This use of synthetic data enables evaluation of the proposed system and explainability pipeline without exposing confidential or personally identifiable information [12].

4.2 Baseline Model Implementation

Based on the transformer-based NLP foundations discussed in Section 2.2, a pretrained transformer-based language model was selected as the baseline ar-

chitecture for document-level information sensitivity classification.

A pretrained transformer model was selected in order to leverage contextual language representations learned during large-scale pretraining while adapting the model to the domain-specific classification task through supervised fine-tuning.

Previous work discussed in Section 2.2.4 demonstrates the suitability of transformer-based architectures for legal and regulatory document classification tasks.

The baseline implementation prioritizes robustness and reproducibility over extensive architectural modification. Avoiding unnecessary changes to the model architecture helps ensure that any observed differences in interpretability, trustworthiness, or performance can be attributed primarily to the explainability components rather than to variations in model complexity.

4.2.1 Model Architecture

The modelarchitecture consists of a pretrained transformer encoder with a lightweight classification head for multi-class prediction. The architecture does not extend or modify the internal structure of the transformer encoder. Instead, a standard fine-tuning approach is employed, where the final hidden representation produced by the transformer, corresponding to a document-level embedding, is passed to a fully connected classification head. This classification head outputs a probability distribution over the predefined information sensitivity classes. Standard fine-tuning has been shown to be effective for document classification tasks, including legal and regulatory document classification and provides a well-established and reliable modelling approach [3, 17].

The decision to employ a simple classification head is motivated by both methodological and practical considerations. Adding more layers or architectural components would increase model complexity and the risk of overfitting, particularly given the limited size and controlled nature of the dataset. More importantly, maintaining a simple architecture is essential for isolating and evaluating the impact of the explainability techniques employed in this study. By minimizing architectural complexity, any observed effects on interpretability or trustworthiness can be more confidently attributed to the explainability methods rather than to changes in model structure.

4.2.2 Training Procedure

The model is trained using a supervised fine-tuning approach, where a pretrained transformer model is adapted to the information sensitivity classification task using labelled documents from the dataset.

Standard optimisation techniques commonly used for transformer-based models

are employed during training. The Adam optimiser [29] was used during training together with weighted categorical cross-entropy loss. Model parameters are updated by minimizing a weighted categorical cross-entropy loss function, which is appropriate for multi-class classification tasks and can emphasise the importance of more secure document types.

To ensure training stability, a fixed number of training epochs and conservative learning rates are used based on prior work and preliminary experimentation [3]. These choices help prevent unstable training dynamics and reduce the risk of overfitting, particularly when working with limited and controlled datasets.

Extensive hyperparameter tuning is deliberately avoided in order to reduce computational costs and to prioritize reproducibility. This design choice ensures that the resulting model configuration can be reliably reproduced and evaluated within real-world and regulated deployment environments. No additional regularization techniques are introduced beyond those already present in the pretrained transformer architecture, such as dropout, unless clear signs of overfitting are observed [3].

4.3 Explainability Integration

Explainability mechanisms were integrated into the classification pipeline in order to support inspection and analysis of model predictions within a regulated financial context. Explainability was treated as a core evaluation component throughout the study. In this study, feature attribution explanations are implemented using Integrated Gradients, while attention-based explanations are derived directly from the transformer's self-attention mechanisms.

The study adopts a post-hoc explainability approach, where explanations are generated without modifying the underlying transformer architecture. Post-hoc explainability allows explanations to be generated for specific prediction results without changing the architecture of the model itself, thereby preserving the prediction accuracy of the transformer architecture-based classification model [12, 23].

Two complementary explanation techniques were implemented: attention-based indicators and feature attribution explanations. These techniques offer different views on how the classification model makes decisions, thereby providing an opportunity to analyse the explanations at different levels, including token-level and document-level representations.

Attention-based indicators are interpreted together with feature attribution results and consistency between these two views is treated as an important signal of explanation reliability.

4.3.1 Attention-Based Explanations

Attention weights were extracted from the transformer model in order to analyse how contextual focus was distributed across input tokens during inference. In this study, attention-based indicators were used as qualitative transparency tools rather than faithful causal explanations.

4.3.2 Feature Attribution Methods

Integrated Gradients was implemented in order to provide token-level feature attribution explanations complementary to the attention-based analysis. In this study, Integrated Gradients is used as the primary feature attribution method [6].

In the current study, feature attribution explanations are provided at the token level, with direct reference to the original text of the input documents. This design choice is intended to ensure that explanations remain interpretable, enabling domain experts and auditors to assess whether the model's predictions are grounded in semantically meaningful and policy-relevant information rather than in irrelevant or potentially spurious cues [20].

Feature attribution methods were evaluated with respect to wether highlighted tokens corresponded to semantically meaningful indicators of information sensitivity. By providing explicit evidence of the input elements that most strongly influence a predicted class label, these methods support compliance assessment and human oversight [26]. In addition, feature attribution explanations are used to triangulate model behaviour in conjunction with attention-based explanations, thereby enhancing confidence in the transparency and trustworthiness of the automated classification system.

4.4 Performance Analysis

The evaluation combines traditional classification metrics with qualitative explainability analysis in order to assess both predictive performance and transparency characteristics. The traditional quality measurement techniques focus on evaluating the system's ability to distinguish between sensitive and non-sensitive data. On the other hand, quality measurements focusing on explanations are used to assess the quality of decision explanations produced by techniques such as attention-based explanations and Integrated Gradients attributions, which is particularly important in regulated and audit-sensitive application contexts.

4.4.1 Classification Metrics

Classification performance was evaluated using accuracy, precision, recall and F1 score derived from the confusion matrix. Since classification errors involving sensitive documents may have different practical consequences depending on

the information sensitivity class, both class-level and aggregate metrics were considered during evaluation.

4.4.2 Explainability Evaluation

Explainability evaluation focused primarily on qualitative inspection of explanation consistency, interpretability and semantic relevance. This study does not attempt to measure trust or auditability as human or organisational constructs. Instead, explainability is evaluated in terms of whether the generated explanations provide the technical affordances required to support transparency, human oversight and audit processes in regulated financial environments.

In this study, explainability evaluation is primarily qualitative in nature and is based on the triangulation of findings from two complementary techniques: attention-based explanations derived from the transformer model and feature attributions computed using Integrated Gradients. The evaluation focuses on whether the explanations generated by these methods highlight linguistically or semantically meaningful content corresponding to known indicators of sensitive information, as well as whether the explanations produced by the two methods are consistent across documents with similar characteristics. Consistency and interpretability are treated as key indicators of explanation reliability in the absence of formal ground-truth explanations [30].

Due to the applied scope and controlled experimental setting of the study, explainability evaluation prioritizes qualitative interpretability analysis over comprehensive quantitative benchmarking.

4.5 Success Criteria

Success criteria were defined across three dimensions: predictive performance, explainability and industrial applicability

4.5.1 Predictive Performance Criteria

Predictive performance is essential for ensuring the viability of an automated document classification system. The model must be able to differentiate between various levels of information sensitivity with sufficient accuracy to reduce reliance on manual classification, while ensuring that the risks associated with misclassification are kept to a minimum [3].

Success criteria for predictive performance are defined based on the model's ability to produce stable and consistent results across standard multi-class classification metrics, including accuracy, precision, recall and F1-score. Particular emphasis is placed on balanced performance across all sensitivity classes, as misclassification of high-sensitivity information carries significantly greater risk than misclassification of lower-sensitivity documents [12].

Due to the controlled and synthetic nature of the dataset used in this research, the performance of the model is not compared against state-of-the-art systems. Instead, performance is evaluated relative to the baseline transformer-based model presented in Section 4.2. Improvements or degradations in predictive performance are assessed in relation to this baseline to ensure that the integration of explainability mechanisms does not lead to an unacceptable reduction in classification performance. A successful outcome is therefore defined as the ability to incorporate explainability mechanisms without materially compromising classification accuracy [20].

4.5.2 Explainable Classification Criteria

The concept of explainability is an important measure of success in this study, as the proposed system is designed for application in a regulated financial setting where decisions are expected to be transparent and traceable through attention-based explanations and Integrated Gradients–based feature attribution. In such environments, it is not sufficient for an automated system to produce accurate predictions; it must also provide explanations that support human understanding, oversight and accountability [12, 20].

Explainability is considered successful if it allows human reviewers to determine which parts of a document contributed to a given classification decision and to evaluate whether these contributions are semantically meaningful and relevant to the information sensitivity classification task. Attention-based explanations and feature attribution methods are considered successful if, through qualitative evaluation, the highlighted parts of the document correspond to plausible and domain-relevant indicators of sensitive information, as would be expected based on organisational and regulatory knowledge [24, 25].

The trustworthiness of the system is supported if the explanations provided are consistent across similar documents and make intuitive sense to human reviewers. Although this study does not conduct a formal evaluation of user trust, explanations are considered successful if they support human reasoning about the system's decision-making process rather than functioning as opaque or potentially misleading artifacts [20].

The application of multiple explanation techniques is also considered a measure of success, as the use of complementary explanation methods is expected to mitigate the limitations of individual techniques and improve confidence in the transparency and reliability of the system's behaviour [12].

4.5.3 Industrial Applicability Criteria

In addition to security level identification criteria, the proposed system must satisfy criteria related to compliance and practical usability in order to be con-

sidered successful within a regulated financial domain. These criteria reflect the conditions under which an automated document classification system is expected to operate in practice.

From a compliance perspective, the system is considered successful based on its compatibility with regulatory and organisational requirements rather than on direct access to sensitive data. The proposed approach should demonstrate that a transformer-based document classification system, combined with explainability mechanisms, can operate in a manner that supports auditability and regulatory review without exposing sensitive information or violating organisational constraints [26].

From a usability perspective, the system is considered successful if it can function as a decision-support tool within document handling workflows in a regulated financial environment. The explanations provided should be understandable to relevant user groups, such as information security specialists or compliance officers and should support manual review and informed decision-making during the document classification process [20].

Success in this dimension is achieved if the proposed methodology demonstrates the feasibility of accurate and explainable document classification under realistic regulatory and organisational constraints. This includes showing that meaningful explanations can be generated and inspected while preserving human oversight and accountability within the document classification process.

5 Implementation

This section documents how the proposed document classification system is constructed, including dataset generation, preprocessing, model development, training and explainability integration. All elements of the implementation are based on the methodological framework described previously, with an emphasis on reproducibility, clarity of purpose and compatibility with a regulated financial environment. The overall implementation is structured into modular components corresponding to data generation, model training and explainability analysis, allowing for controlled experimentation and systematic evaluation. Particular focus is placed on maintaining a stable baseline model and a clear processing pipeline, ensuring that any observed results can be attributed to the applied methods rather than to differences in implementation across experiments.

5.1 Dataset Construction

The current experiment involves a dataset that is synthetically generated using a predefined template-based approach. This strategy is adopted due to the limited availability of real-world financial documents, while still enabling controlled experimentation within a regulated context.

The dataset consists of four classes of text, each representing a distinct level of information sensitivity according to standard classification schemes used in financial institutions. Furthermore, the dataset is constructed as a balanced multi-class classification dataset, where each class contains an equal number of documents.

Documents are generated using a label-conditioned template-based approach, where each class is associated with a set of archetypal document structures and content patterns. These archetypes are designed to reflect realistic document categories, including internal reports, financial summaries and client communications. For each document, an appropriate template is selected based on the assigned label and populated with dynamically generated content.

The generation process incorporates several randomised elements to introduce variability within the dataset. Attributes such as department names, dates, numerical values (e.g., financial figures or budget allocations) and domain-specific contextual phrases are randomly generated for each document. This ensures that documents within the same class share underlying structural and semantic characteristics while maintaining sufficient diversity.

Finally, the dataset is shuffled to eliminate potential ordering effects during training and evaluation. Each instance consists of a text-label pair, where the label represents the assigned information sensitivity level. This structure enables the classification model to learn linguistic and contextual patterns associated

with different sensitivity classes.

Overall, the dataset supports controlled experimentation and enables the evaluation of the proposed system with respect to both classification performance and explainability. This design balances the need for realistic document representation with the constraints imposed by data privacy and regulatory compliance.

5.2 Preprocessing Pipeline

Before training the model, documents in the dataset undergo a series of preprocessing steps aimed at preparing the data for compatibility with the transformer-based architecture.

Given that the dataset is synthetically generated, the text does not require extensive preprocessing, as it is produced in a controlled and well-structured format.

First, tokenization is applied using the `BertTokenizer` associated with the pretrained `bert-base-uncased` model. Each document is converted into a sequence of token identifiers, along with corresponding attention masks required by the transformer architecture.

Moreover, each token sequence is automatically padded and truncated where necessary to meet the fixed input size constraints of the model. This ensures that all sequences have a consistent length, enabling efficient batch processing during training [19].

No extensive preprocessing techniques, such as stopword removal or stemming, are applied, as transformer-based models are designed to operate effectively on raw text inputs and preserve contextual information [3].

The dataset is then divided into two subsets: a training set and a validation set. This division is performed using a stratified sampling approach to ensure that the class distribution is preserved across both subsets.

Furthermore, during preprocessing, a direct correspondence between the original text and its tokenized representation is preserved. This property is essential for the application of explainability techniques, as it enables attribution scores and attention patterns to be traced back to the corresponding tokens in the original documents [12, 23].

5.3 Model Implementation

The document classification model is implemented using a transformer-based architecture, a pretrained BERT model. In accordance with the implementation framework specified in Section 4.2, the pretrained model is built using the

`transformers` module in combination with the PyTorch deep learning library.

A `bert-base-uncased` model is used as a basis for the model development. The selected pretrained model was trained on large-scale general-domain texts and can be effectively fine-tuned with labelled data from a specific domain [3]. The choice of this particular model fits the methodological approach mentioned in Section 4.2, in which transfer learning is performed to address issues related to the small size and synthetic nature of the dataset.

For the implementation of the classification problem, the pretrained BERT model is customized with the addition of a classification layer. Specifically, the `BertForSequenceClassification` class is used, which builds a classifier on top of a pretrained transformer-based encoder. The added classification layer includes one fully connected layer applied to the context embedding of the input sequence. The number of classes in the output layer is set to 4, as there are four information sensitivity categories in the input data: Public, Internal, Confidential and Restricted. The model is configured for single-label multi-class classification, ensuring compatibility with the categorical nature of the task.

The model accepts tokenized input texts produced during preprocessing, including token identifiers and attention masks.

In order to enable efficient batch processing during the training phase, a dataset class is created. The implemented dataset encapsulates tokenized inputs and their corresponding labels, enabling efficient batching and integration with the training pipeline.

The process of training and evaluating the model will be handled via HuggingFace's Trainer API, which offers a high level of functionality regarding the fine-tuning process itself, allowing the implementation of the training loop with a sufficient amount of flexibility and reliability. The Trainer API ensures consistent handling of batching, gradient updates and evaluation procedures, minimizing the risk of implementation errors.

There are no changes introduced in terms of the architecture of the pretrained transformer model. This choice of design reflects the methodology goal of having an experiment based on a stable and well-known baseline model. This decision minimizes risks related to architectural changes in the model and their influence on its behaviour. This also ensures that any observed effects on interpretability can be attributed to the explainability techniques rather than to changes in model architecture. It also makes it less prone to overfitting [31].

In general, the current implementation strategy represents a standard fine-tuning approach for transformer-based document classification tasks [3, 17].

5.4 Training Configuration

The training procedure for the document classification model is configured to ensure reliable fine-tuning of the pretrained transformer while satisfying reproducibility requirements and methodological restrictions stated in Section 4.2.

The model training procedure is based on the supervised fine-tuning approach applied to the synthetically generated dataset mentioned in Section 4.1. The data is split into training and validation sets through a stratified sampling approach, ensuring that 80% of the data belongs to the training set while 20% is included in the validation set. The stratified sampling approach helps to ensure that the same class distribution is represented in both datasets, especially for multi-class classification problems with different sensitivity levels [32]. Reproducibility is ensured through the use of a fixed random seed when splitting data into training and validation subsets.

The training procedure is performed for a fixed number of five epochs. Such an approach helps to ensure that the model learns effectively from the data without risking overfitting due to an excessive number of epochs, since the data size is limited and artificially created. The model training procedure is carried out using mini-batch gradient descent with the batch size being equal to eight samples.

The training process itself is managed by the AdamW optimiser [29], which is the default option for training transformers. AdamW has been proven to be suitable for fine-tuning pretrained BERT models, helping achieve effective results. The learning rate for training is set at $5 * 10^{-5}$, which corresponds to standard practices for fine-tuning transformer models [3]. Moreover, weight decay with a coefficient equal to 0.01 is used to reduce the risk of overfitting [31].

A learning rate warmup technique is used for stabilization purposes, where the rate is increased gradually in the first stages of training [3, 19]. In particular, a warmup ratio equal to 0.1 is used, meaning that 10% of the total training steps are used for warmup before reaching the maximum learning rate. In this way, it is possible to reduce destabilizing effects of large gradient updates made by the model during the initial phase of training.

Categorical cross-entropy loss is used for training the model in the problem under consideration. In order to adjust class importance, a weighted version of the loss function is adopted for the multi-class classification task. In this way, greater emphasis is placed on sensitive classes, introducing a higher penalty for incorrect classifications of information that falls within sensitive categories. This solution is adopted in accordance with practical requirements of information security document classification problems.

Model evaluation is conducted at the end of each epoch. Classification-related evaluation metrics are used for determining the performance of the model, with

accuracy and weighted F1 score being considered. The weighted F1 score is chosen as the main evaluation metric, as it takes into account both precision and recall, while also reflecting the class distribution.

In order to select the best-performing model, a model checkpointing technique is applied. According to this technique, the model with the highest value of weighted F1 score on the validation set is automatically selected as the final model, thus taking into account generalization capability rather than performance only at the final training stage.

In summary, the proposed training configuration follows a conventional fine-tuning procedure for transformer-based models, while incorporating specific adaptations in terms of loss function weighting and model selection criteria in order to achieve satisfactory results.

5.5 Explainability Implementation

The implementation of explainability within the document classification system is designed to enable understanding of how the underlying transformer model makes predictions, without modifying the architecture of the transformer used by the system. For this reason, a post-hoc explainability approach is employed to examine both attention-based indicators and feature attribution using the Integrated Gradients method [12], as previously described in Section 4.3.

5.5.1 Overview of Explainability Approach

The analysis of the XAI component is conducted in a separate pipeline after the classification model has been trained and is therefore external to the model itself. The purpose of this analysis is to provide explanations of which features of the input text contributed to the model's prediction of a class label.

To accomplish this, two techniques are used: attention-based explanations and feature attribution explanations. The attention-based technique is derived directly from the transformer model, while the feature attribution technique is based on Integrated Gradients, as introduced in prior work.

The use of these two techniques provides complementary perspectives on how the model produces its predictions. Furthermore, it allows for comparison between the explanations generated by each method, thereby strengthening the validity of the interpretation.

5.5.2 Attention-Based Explanation Extraction

Attention-based explanations are obtained by enabling the output of attention weights during model inference. This is achieved by configuring the transformer model with the `output_attentions=True` parameter, which allows access to

attention distributions across all layers and attention heads.

For each input document, attention weights from the final transformer layer are extracted and averaged across all attention heads to produce a single interpretable representation. The resulting attention matrix represents the relationships between tokens in the input sequence.

To generate a document-level explanation, the attention weights associated with the classification token ([CLS]) are examined. The [CLS] token serves as a composite representation of the entire input sequence in BERT-based classification tasks [3]. By analysing the distribution of attention from the [CLS] token to all other tokens, it is possible to identify which parts of the input the model focuses on when producing a classification.

Finally, the attention scores are mapped back to the corresponding tokens in the original text, enabling visual interpretation of the model's focus during the document classification task.

5.5.3 Integrated Gradients Attribution

Integrated Gradients is used as the feature attribution method to complement the attention-based explanation approach. As a gradient-based attribution technique, Integrated Gradients assigns an importance score to each input feature based on its contribution to the model's prediction [6].

The Captum library is used to implement the Integrated Gradients method within the PyTorch framework. For each input document, attribution scores are computed at the token level with respect to a selected target class, typically corresponding to the predicted label of the model.

Attribution is determined by calculating gradients along a defined path from a baseline input, representing a neutral reference, to the actual input, defined as the tokenized document. The resulting attribution scores reflect the contribution of each token to the model output for the specified target class.

The attribution scores are aggregated at the token level and aligned with the original text of the input document, enabling direct interpretation of which words or phrases most strongly influence the model's classification decision.

5.5.4 Token-Level Alignment and Filtering

To maintain interpretability and utility of explanations, a direct mapping between tokenized inputs and the original text is preserved throughout the preprocessing and explainability pipeline.

Special tokens introduced by the transformer model for classification ([CLS]),

separation ([SEP]) and padding ([PAD]) are excluded from the explanation analysis. These tokens do not represent meaningful linguistic content and may otherwise distort the interpretation of attribution scores or attention visualizations [23].

By filtering out these tokens, the resulting explanations focus exclusively on semantically relevant components of the input text, thereby improving clarity and interpretability for human reviewers.

5.5.5 Selection of Representative Examples

Explainability analysis is performed on a subset of validation documents that represent different classification scenarios:

- Correctly classified documents from low-sensitivity classes (e.g., Public)
- Correctly classified documents from high-sensitivity classes (e.g., Restricted)
- Misclassified documents where the predicted label does not match the ground truth label

This selection strategy allows for comparison of model behaviour across both correct and incorrect predictions. It enables evaluation of the strengths and weaknesses of the model's reasoning, particularly in cases where misclassification may be influenced by misleading or inappropriate textual cues.

5.5.6 Comparative Analysis of Explanation Methods

Both attention-based explanations and Integrated Gradients attributions are used as complementary approaches to assess consistency and interpretability. Consistency between these two methods is treated as a proxy for reliability in the explanation of the model [20, 26, 33].

If both attribution scores and attention weights highlight similar regions of the input text, this increases confidence that those regions are genuinely influential in the model's decision-making process. Conversely, if the two methods emphasise different regions, these discrepancies are examined to identify potential limitations in the explainability techniques or ambiguities in the model's internal representations.

By combining both methods, the approach provides a more comprehensive understanding of model behaviour, supporting the requirements of transparency and auditability in regulated financial environments.

5.6 Implementation Tools and Environment

Python 3.11 was used to develop the proposed document classification system, using widely adopted libraries for natural language processing, deep learning

and model interpretability. All development and experimentation were conducted in Visual Studio Code, which served as an integrated environment for code development, execution and debugging.

The system is implemented using the PyTorch deep learning framework, which enables efficient tensor computations and supports automatic differentiation. Building on PyTorch, the HuggingFace Transformers library is used to access and fine-tune a pretrained BERT model (`bert-base-uncased`) employed in this study. The library provides both the model architecture and its corresponding tokenizer, ensuring consistency between text preprocessing and model input representation.

Model training and evaluation are managed using the HuggingFace Trainer API. This API abstracts key components of the training loop, such as batching, optimisation and evaluation, enabling reliable and reproducible fine-tuning. In addition, it simplifies the configuration and management of training parameters.

Standard Python libraries are used throughout the preprocessing pipeline and for dataset handling. These libraries support data manipulation and numerical computation tasks required during dataset construction, including stratified splitting into training and validation sets, as well as the preparation of inputs for model training.

Explainability analysis is performed using the Integrated Gradients method, implemented via the Captum library. Captum is designed to provide interpretability for PyTorch models and enables the computation of feature attribution scores at the token level. This allows explainability techniques to be applied directly to the trained classification model without requiring architectural modifications.

The implementation is organised into separate components corresponding to dataset construction, model training and explainability analysis. This modular design improves clarity and reproducibility and aligns with the methodological separation between data generation, model development and evaluation.

To ensure reproducibility, fixed random seeds are applied during dataset splitting and model training. This ensures that results obtained under the same configuration remain consistent across multiple runs.

Overall, the selected tools, environment and implementation structure provide a consistent and reproducible framework for developing transformer-based document classification systems and integrating explainability techniques within a controlled experimental setting.

6 Results and Analysis

This section provides a detailed analysis of the experimental results obtained from the proposed transformer-based document classification model, together with an assessment of its predictive performance and explainability characteristics. The evaluation focuses not only on the model's ability to correctly classify documents according to the predefined information sensitivity levels, but more importantly on how the model reaches its decisions and whether those decisions are interpretable within the context of regulated financial environments.

The evaluation is divided into several components. The first component presents commonly used classification metrics that assess predictive performance across the predefined security classes. These metrics include accuracy, precision, recall, F1 score and confusion matrix analysis, presented in Sections 6.1.1, 6.1.2 and 6.1.3. In addition, selected prediction examples are presented in Sections 6.2.1 and 6.2.2 in order to provide further insight into the model's behaviour at the document level.

This section also evaluates explanation outputs generated using attention-based indicators and Integrated Gradients feature attribution methods. These analyses aim to identify which linguistic and contextual elements contribute most strongly to the model's predictions, as well as whether the explanations produced by the different explainability techniques are consistent and interpretable. Relationships between attention distributions and feature attribution scores, including overlap and consistency comparisons across multiple examples and document classes, are also examined.

In addition to assessing prediction accuracy and local explanation outputs, the evaluation also examines model confidence levels and measures predictive uncertainty through entropy analysis in cases where classifications may be ambiguous. High predictive accuracy alone does not necessarily indicate that a model is robust or trustworthy. Therefore, additional analyses are conducted to investigate the presence of shortcut learning, such as reliance on repeated lexical or structural patterns rather than contextual understanding, as well as potential token leakage within the dataset.

Finally, the results are interpreted in relation to the objectives of the study concerning transparency, auditability and explainability in regulated financial environments. Explanations are not treated solely as visualisation components, but are instead evaluated according to whether they provide meaningful support for human oversight and compliance-oriented review processes.

6.1 Classification Performance

The transformer-based model for document classification was evaluated using standard multi-class classification metrics, including accuracy, precision, recall

and F1 score. Evaluation was conducted on the validation subset of the synthetically generated dataset described in Section 5.1, which contains an equal number of documents across the four information sensitivity levels: Public, Internal, Confidential and Strictly Confidential.

The validation dataset contains a total of 320 samples, with 80 samples belonging to each class.

6.1.1 Overall Performance

The model achieves perfect performance across all evaluation metrics:

Metric	Score
Accuracy	1.000
Precision (macro)	1.000
Precision (weighted)	1.000
Recall (macro)	1.000
Recall (weighted)	1.000
F1 Score (macro)	1.000
F1 Score (weighted)	1.000

These results show that the model correctly classifies all validation samples with no errors.

The absence of classification errors suggests that the model is highly effective at identifying patterns associated with the defined sensitivity levels within the controlled dataset environment. However, the perfect performance also indicates that the synthetic of the dataset may not contain sufficient variability or ambiguity to fully challenge the model. The strong separability also likely simplifies the classification task compared to real-world environments.

6.1.2 Per-Class Performance

Performance at the class level is also uniformly perfect. For each of the four classes (Public, Internal, Confidential and Strictly Confidential), the model achieves:

Metric	Score
Precision	1.000
Recall	1.000
F1 Score	1.000

Each class is represented by 80 validation samples and no instances are misclassified. This indicates that the model is able to completely separate the classes within the validation dataset.

Although performance is identical across all classes, some differences can still

be observed in the confidence distribution by the model. Internal documents achieve the highest average confidence (0.9985) together with the lowest entropy values, while Public documents exhibit slightly lower confidence (0.997) and higher entropy compared to remaining classes, suggesting higher linguistic overlap among lower-sensitivity documents.

6.1.3 Confusion Matrix and Error Analysis

No misclassified instances are observed in the validation results. As a result, traditional error analysis based on validation set misclassifications is limited in this case. Figure 1 illustrates the confusion matrix for the classification model.

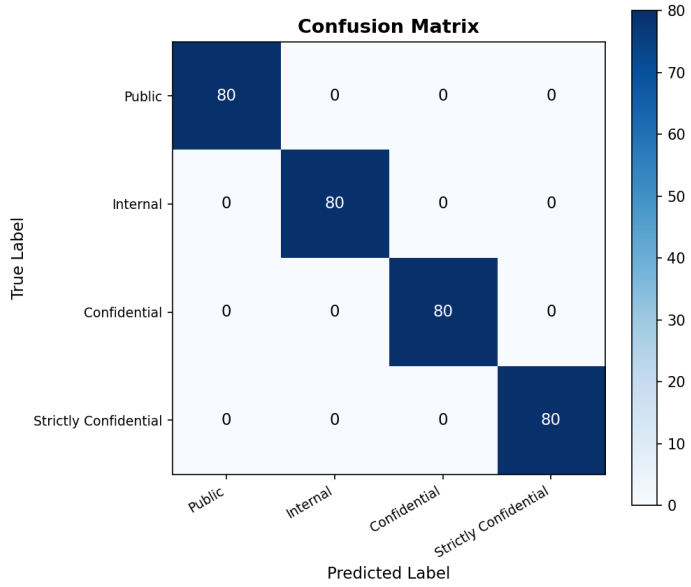

Figure 1: Confusion matrix for the document classification model evaluated on the validation dataset.

The absence of errors indicates that the validation dataset and by extension the entire dataset, is highly separable. However, perfect validation performance does not necessarily imply robust generalization to more realistic or less controlled environments.

Additional robustness analysis supports this. External synthetic evaluation using example documents that deviate from the structured patterns introduce misclassifications. For example:

- A document labeled as **Public** is classified as **Confidential**

- A document labeled as **Internal** is classified as **Strictly Confidential**

These cases demonstrate that the model partially relies on recurring lexical and structural patterns within the synthetic dataset. This interpretation is further supported by the token leakage and shortcut learning analysis presented in section 6.5.

The absence of misclassifications within the validation set therefore mostly reflects the controlled nature of the dataset design. This suggests that the validation results should primarily be interpreted as evidence that the model successfully learned the synthetic dataset. While this enables stable evaluation of explainability techniques, it also limits conclusions regarding real-world application.

Because no validation misclassifications were available for detailed inspection, the explainability analysis primarily focuses on correctly classified representative examples across all classes.

6.1.4 Summary

The classification model demonstrates perfect predictive performance on the validation dataset, achieving complete accuracy across all evaluation metrics and document classes.

The results indicate that transformer-based language models are highly effective at recognizing linguistic and structural patterns in a controlled financial document environment. However, further robustness analysis reveals that the observed performance is tied closely to the structured nature of the dataset.

These findings highlight the importance of interpreting classification metrics in relation to dataset design. Although the model demonstrates strong capability within the experimental setting, further evaluation using more diverse and less controlled financial documents would be necessary in order to assess real-world generalization and operational robustness.

Therefore, the reported performance metrics should not be interpreted as an upper bound for real-world financial document classification performance. Instead, they should demonstrate the effectiveness under controlled conditions, intended to provide support for analysis of explainability.

6.2 Explainability Results

This section presents the results of the explainability analysis applied to the transformer-based document classification model. The focus is on examining how the model arrives at its predictions, specifically by analysing the contribution of individual tokens (words) to the classification decisions. Two complementary techniques are used for this purpose: attention-based indicators (Sec-

tion 6.2.1) and Integrated Gradients (IG) for feature attribution (Section 6.2.2).

Despite the perfect classification performance observed in Section 6.1, explainability is the most critical component of the evaluation. High predictive accuracy does not guarantee that the model relies on meaningful or robust features. Additionally, the use of synthetic and controlled data means that the model may exploit superficial or class-specific patterns. Therefore, the analysis in this section focuses on understanding what the model has learned rather than how well it performs.

The results are illustrated using representative examples of correctly classified documents across different sensitivity levels. For each example, both techniques identify the most influential tokens contributing to the model's prediction. This provides qualitative insight into the model's behaviour and supports evaluation of its interpretability and transparency.

In addition to analysing the individual explanation methods, this section also examines the consistency between them. Consistency between these methods is treated as an indicator of explanation reliability, while divergence suggests the model relies on different internal signals that are not directly aligned.

6.2.1 Attention-Based Explanations

The attention-based explanations were analysed across multiple correctly classified examples from each sensitivity class. Given the perfect classification performance on the validation set, the analysis focuses on identifying recurring attention patterns and assessing wether they align with semantically meaningful features related to information sensitivity.

The self-attention mechanism in transformer models enables the extraction of attention-based explanations by assigning weights to tokens when constructing a document-level representation. These attention weights indicate which tokens the model focuses on when producing a classification decision.

Attention weights were frequently concentrated on a combination of domain-specific terms (e.g., bank, financial, policy) and structural or contextual tokens (e.g., organisation names, document headers). This suggests that the model leverages both semantic content and document structure when forming predictions.

For Public documents, attention is often docused toward general business and communication-related terms, such as references to public-facing activities, financial summaries and organisational identifiers. Contrasting this, Strictly Confidential documents show increased attention on terms related to governance, finance, regulatory processes and internal decision-making.

In addition to these extreme cases, the mid-level sensitivity categories also exhibit distinct patterns. For Internal documents, attention is often distributed across operational and organisational terms, such as references to policies, systems, teams and internal processes. This reflecting internal activities and and governance, which differentiate them from both public-facing and high sensitivity documents.

For Confidential documents, attention tends to focus on terms associated with sensitive business operations, such as procurement, compliance, financial negotiations, or personnel-related information. Compared to Internal documents, these attention patterns reflect a shift toward more sensitive and decision-critical content, though not to the same extent as Strictly Confidential documents.

Below are some examples of influential tokens for the different classes:

Public Document Example For a correctly classified Public document, attention is concentrated on tokens such as:

public	*published*	*distribution*	*samplebank.com*

These tokens are directly related to the document's distribution and accessibility, which are strong indicators of low sensitivity. In particular, the presence of terms such as "published" and references to public domains suggests that the document is intended for external dissemination.

The attention distribution shows that the model prioritizes explicit signals of public availability rather than unrelated contextual information. Lower attention weights assigned to tokens such as "mobile" or "transaction" further indicate that domain-specific content is less influential than distribution-related cues in determining sensitivity.

Internal Document Example For a correctly classified Internal document, attention is concentrated on tokens such as:

policy	*systems*	*teams*	*management*	*checkpoint*

These tokens are associated with internal operational processes, governance and organisational coordination. In particular, references to policies, systems maintenance and team-related processes indicate that the document concerns internal workflows rather than externally distributed or highly restricted information.

The attention distribution suggests that the model relies on organisational and procedural terminology when identifying Internal documents. Tokens such as systems, management and checkpoint reflect operational planning and coordination activities that are typically accessible only within the organisation but

do not necessarily contain highly sensitive or strategically critical information.

Compared to Public documents, the attention is less focused on distribution-related language and instead emphasises internal operational structure. At the same time, the absence of highly sensitive regulatory or governance-related terminology differentiates these documents from the Strictly Confidential category.

Confidential Document Example For a correctly classified Confidential document, attention is concentrated on tokens such as:

salary	*procurement*	*consulting*	*HR*	*management*

These tokens are associated with sensitive business activities and restricted organisational processes. In particular, references to procurement negotiations, compensation structures and HR-related activities indicate that the document contains information intended for limited internal access.

The attention distribution demonstrates that the model identifies confidentiality through terms linked to business negotiations, personnel matters and financial decision-making. Tokens such as "salary" and "procurement" provide strong indicators of commercially or organisationally sensitive content.

Compared to Internal documents, the attention patterns in Confidential documents are more associated with restricted business operations and decision-critical activities. However, the attention remains less focused on executive governance and regulatory terminology than in Strictly Confidential documents.

Strictly Confidential Document Example For a correctly classified Strictly Confidential document, attention is concentrated on tokens such as

insider	*directors*	*only*	*board*	*distribution*

These tokens indicate restricted access and are associated with internal governance and regulatory structures, suggesting that the document is highly sensitive. Several of the highly attended tokens appear together in contextual expressions such as "Board of Directors only", which indicates limited distribution and strong confidentiality constraints. This suggests that the model is not relying solely on isolated keywords such as *only*, but rather on combinations of semantically related tokens that together signal high information sensitivity.

In addition, attention is assigned to tokens such as *information* and *constitutes*, indicating that the model may also consider contextual language related to formal definitions and regulatory framing.

Observations Across both examples, attention-based explanations show that the model consistently prioritizes:

- Distribution indicators (e.g., public vs. restricted access)
- organisational roles (e.g., board, directors)
- Sensitivity-related terminology

This suggests that the model relies on semantically meaningful indicators that align with common information classification policies. However, attention distributions also revealed limitations. In several cases, high attention weights were assigned to frequently occurring or structurally common tokens (e.g., organisation names or repeated terms), which may not directly contribute to the semantic distinction between sensitivity classes. This indicates that attention alone may not provide a reliable explanation of model behaviour.

These observations motivate the use of complementary explanation methods, such as Integrated Gradients, which provide a more direct measure of feature contribution. The following section examines whether attribution-based explanations offer more precise insights into the model's decision-making process.

6.2.2 Integrated Gradients Attributions

Integrated Gradients (IG) was applied to generate token-level feature attributions for the same set of correctly classified documents analysed in the previous section. Given the perfect classification performance observed on the validation set, the purpose of this analysis is not to distinguish between correct and incorrect predictions, but to examine which input features contribute most strongly to the model's decisions across different sensitivity classes.

To complement the attention-based analysis, IG is used to quantify the contribution of individual tokens to the model's predictions. Unlike attention-based methods, IG provides a gradient-based measure of feature importance, offering a more direct indication of how input tokens influence the model's output.

In contrast to attention-based explanations, Integrated Gradients provides a more direct measure of feature importance by quantifying the contribution of each token to the final prediction. The resulting attribution scores highlight tokens that have a strong positive or negative influence on the model's decision, offering a more precise view of the underlying decision-making process. Across all sensitivity classes, IG attributions reveal clearer and more semantically meaningful patterns compared to attention weights.

For Public documents, high-attribution tokens are primarily associated with general business communication and public-facing language, such as references to customers, products, financial performance and organisational activities.

For Internal documents, IG highlights tokens related to operational processes and internal coordination, including terms such as systems, management, updates and internal procedures. These attributions reflect internal activities and organisational workflows.

For Confidential documents, attribution scores are concentrated on tokens associated with sensitive business operations, including procurement, compliance, negotiations and personnel-related terms. These attributions indicate that the model identifies content linked to restricted business processes and decision-making contexts.

For Strictly Confidential documents, IG emphasises highly sensitive and context-specific tokens, such as references to financial supervision, regulatory authorities, strategic transactions and explicitly sensitive terminology (e.g., confidential). These attributions suggest that the model relies on strong semantic indicators of high sensitivity.

Public Document Example For the same *Public* document, the highest attribution scores are assigned to:

public	*published*	*distribution*	*on*	*com*

These tokens largely align with those identified by the attention-based method. Notably, *public* and *published* receive significantly higher attribution scores than the other tokens, suggesting that they are the primary drivers of the classification decision.

Some tokens exhibit negative attribution scores, such as the author name (*Strand*). This indicates that these tokens may either reduce the model's confidence in the predicted class or are not relevant for determining the classification.

Internal Document Example For the same Internal document, the highest attribution scores are assigned to:

teams	*systems*	*management*	*schedules*	*milestone*

These tokens are strongly associated with operational coordination and internal processes. The high attribution scores assigned to systems and management indicate that the model relies heavily on workflow-related terminology when identifying Internal documents.

Unlike attention-based explanations, IG highlights tokens that are more directly connected to procedural activities and project coordination. Terms such as scheduled and milestone suggest that the model associates Internal documents with planning, maintenance and organisational oversight activities.

Some numerical tokens also receive attribution scores, likely reflecting structural patterns within the synthetic dataset rather than semantically meaningful indicators. This suggests that while IG provides more precise explanations than attention, some attribution patterns may still be influenced by formatting or template-related artifacts.

Confidential Document Example For the Confidential document, the highest attribution scores are assigned to:

compensation	*salary*	*comittee*	*bid*	*tender*

These tokens are directly related to sensitive business operations and personnel-related information. In particular, the high attribution assigned to "compensation" and "salary" suggests that the model strongly associates financial and HR-related terminology with the Confidential sensitivity level.

The attribution scores further indicate that the model relies on procurement and negotiation-related language when identifying Confidential documents. Tokens such as "bid" and "tender" reflect restricted commercial processes and ongoing business negotiations that are not intended for public or broad internal distribution.

Compared to attention-based explanations, IG provides more semantically focused signals for the Confidential category. The highlighted tokens correspond closely to organisationally sensitive activities and align well with expected confidentiality indicators within regulated financial environments.

Strictly Confidential Document Example For the *Strictly Confidential* document, the highest attribution scores are assigned to:

insider	*information*	*sensitivity*	*disclosure*	*regulations*

These tokens are strongly associated with regulatory language and confidentiality requirements. In particular, the high attribution assigned to *insider* reflects the model's sensitivity to terminology commonly linked to restricted or legally protected information.

Observations Integrated Gradients reveals that:

- The model relies heavily on explicit sensitivity-related keywords
- Regulatory and compliance-related language plays a significant role

Furthermore, the alignment between high-attribution tokens and domain-relevant terminology supports the plausibility of the model's reasoning process.

Compared to attention-based explanations, IG attributions tend to emphasise more distinctive and class-specific tokens. For example, tokens such as compensation, salary, breach and compliance receive high attribution scores in Confidential documents, while tokens such as financial, supervisory and authority receive high attribution scores in Strictly Confidential documents. These tokens are more directly aligned with the semantic characteristics of the respective sensitivity levels.

However, IG attributions also exhibit certain limitations. In some cases, high attribution scores are assigned to tokens that are not semantically meaningful, such as punctuation, numerical values, or common function words. This may be due to the controlled nature of the dataset structure or artifacts introduced during preprocessing. Therefore, while IG provides more precise signals than attention, its outputs still require careful interpretation.

The structured nature of the dataset likely contributes to the clarity of the observed attribution patterns. By reducing variability and using structured document templates, the dataset encourages the model to rely on consistent lexical cues associated with each sensitivity class. While this improves interpretability in the experimental setting, it also raises questions regarding the extent to which these attribution patterns would generalize to more diverse real-world data.

Overall, Integrated Gradients provides a more detailed and semantically grounded view of the model's decision-making process compared to attention-based explanations. However, differences between the two methods remain evident. The following section therefore compares attention-based explanations and IG attributions directly, with a focus on their consistency and complementary strengths.

6.2.3 Comparison of Explanation Methods

The previous sections analysed attention-based explanations and IG feature attributions separately. While both methods aim to provide insight into the model's decision-making process, they exhibit notable differences in the type of information they emphasise and the consistency of the resulting explanations.

Overall, attention-based explanations tend to highlight broader contextual and structural patterns within the documents, whereas Integrated Gradients produces more focused and semantically specific feature attributions. Across multiple examples and sensitivity classes, IG consistently identified tokens that were more directly related to the semantic meaning of the corresponding classification category.

For Public documents, both methods identified tokens associated with public communication and external distribution. However, attention weights frequently emphasised repeated structural tokens such as organisation names or

common document markers, while IG assigned higher attribution scores to semantically meaningful terms such as publicly, transparency, sustainability and customers. This suggests that IG provides more direct insight into the content-based reasoning behind the prediction.

A similar pattern is observed for Internal documents. Attention-based explanations often focused on organisational or procedural tokens, including policy, governance and teams. In contrast, IG highlighted operational and workflow-related terms such as systems, scheduled, milestone and maintenance, which more clearly reflect internal organisational processes.

For Confidential and Strictly Confidential documents, the differences between the methods become more pronounced. Attention-based explanations frequently assigned importance to repeated structural tokens and metadata-related terms, while IG emphasised semantically sensitive terminology such as compensation, salary, compliance, disclosure, supervisory and confidential. These tokens are more directly aligned with restricted business operations, regulatory processes and sensitive organisational activities.

The consistency analysis further highlights these differences. Across the evaluated samples, the overlap between the top attention tokens and the top IG attribution tokens remained relatively low. The average Jaccard similarity across all analysed samples was 0.108, with several examples exhibiting no overlap between the methods. In many cases, only one shared token was observed among the top-ranked features.

This low overlap suggests that attention mechanisms and gradient-based attributions capture different aspects of the model's internal behaviour. Attention-based explanations appear to reflect broader contextual relationships and document structure, while IG provides a more localized estimate of token-level contribution to the final prediction.

Despite these differences, both methods reveal partially complementary insights into the classification process. Attention-based explanations provide a high-level overview of contextual focus, while IG offers a more direct indication of which tokens contribute most strongly to the predicted sensitivity level. The combined use of both methods therefore improves interpretability by enabling analysis from multiple perspectives.

At the same time, the observed divergence between the explanation methods also highlights an important limitation of explainability techniques. Since different methods may attribute importance to different features, explanations should not be interpreted as definitive representations of model reasoning. Instead, they provide approximations of the internal decision-making process, each with their own assumptions and limitations.

6.2.4 Class-Specific Explanation Patterns

Analysis of the explanation outputs across different sensitivity classes reveals recurring class-specific patterns in both attention-based explanations and Integrated Gradients attributions. These patterns provide insight into the lexical and contextual signals the model associates with each security classification level.

For Public documents, explanations consistently emphasise tokens associated with external communication, accessibility and publicly available information. Frequently highlighted terms include public, published, website, customers, distribution and sustainability. Both explanation methods indicate that the model strongly associates public-facing language and references to external dissemination with low sensitivity classifications. In many cases, the explanations also focus on organisational identifiers and communication-related terminology, suggesting that the model relies on signals commonly found in externally distributed corporate documents.

Internal documents exhibit a different set of explanation patterns. The model frequently assigns importance to operational and organisational terminology such as systems, teams, policy, governance, maintenance, checkpoint and management. These terms reflect internal coordination, workflow management and procedural activities. Compared to Public documents, the explanations place less emphasis on external distribution indicators and instead focus on language associated with internal organisational processes and collaboration.

For Confidential documents, the explanations become increasingly associated with commercially sensitive and decision-related terminology. Frequently highlighted tokens include *procurement*, *compensation*, *salary*, *consulting*, *HR*, *bid*, *tender* and *negotiations*. These patterns indicate that the model identifies confidentiality primarily through references to restricted business operations, personnel-related information and internal financial activities. The explanations further suggest that the model differentiates Confidential documents from Internal documents by assigning greater importance to terms linked to strategic or commercially sensitive processes.

Strictly Confidential documents demonstrate the most distinct explanation patterns across all classes. Both attention and Integrated Gradients consistently emphasise tokens associated with governance, regulatory oversight, insider information and restricted access. Frequently highlighted terms include *insider*, *board*, *directors*, *supervisory*, *disclosure*, *regulations* and *confidential*. These explanation patterns indicate that the model strongly associates high sensitivity classifications with regulatory language, executive governance structures and explicitly restricted terminology.

Across all classes, the explanation outputs reveal that the model relies heavily

on lexical cues that align with expected sensitivity characteristics. In particular, terms related to access control, organisational hierarchy, financial operations and regulatory processes consistently contribute to classification decisions. This suggests that the model has learned semantically meaningful associations between document content and security classification levels.

At the same time, some explanation patterns also reveal potential limitations in the learned representations. Certain structurally common or repetitive tokens receive importance scores despite having limited semantic relevance to sensitivity classification. This indicates that the model may partially rely on dataset-specific artifacts or recurring template structures in addition to genuinely meaningful semantic features.

Overall, the observed class-specific explanation patterns support the plausibility of the model's reasoning process while also highlighting the influence of structured dataset design on the resulting explanations.

Table 2 summarizes the recurring explanation patterns observed across the different sensitivity classes.

Sensitivity Class	**Frequently Highlighted Tokens**	**Interpretation**
Public	public, published, customers, distribution	Associated with external communication, public availability and publicly distributed organisational information.
Internal	systems, policy, management, teams, checkpoint	Associated with operational workflows, internal coordination, governance and organisational processes.
Confidential	salary, procurement, HR, compensation, tender	Associated with sensitive business operations, personnel-related information and restricted commercial activities.
Strictly Confidential	insider, board, disclosure, regulations, supervisory	Associated with governance structures, regulatory oversight, insider information and highly restricted content.

Table 2: Class-specific explanation patterns observed across explanation methods

6.2.5 Consistency Observations

Consistency between explanation methods was analysed to assess the stability and reliability of the generated explanations across different sensitivity classes and document examples. In this context, consistency refers both to the extent to which different explainability methods highlight similar features and to whether similar documents produce comparable explanation patterns.

The comparison between attention-based explanations and Integrated Gradients (IG) attributions revealed that the overlap between the two methods remained relatively limited across most evaluated samples. In many cases, the methods emphasised different subsets of tokens, even when the predicted class was identical. Quantitative analysis using Jaccard similarity demonstrated generally low overlap between the top-ranked tokens identified by the two methods.

These observations suggest that attention mechanisms and gradient-based attribution methods capture different aspects of the model's internal behaviour. Attention-based explanations often emphasise broader contextual relationships and structural information, whereas IG tends to identify more localized and semantically meaningful contributions to the final prediction.

Despite the relatively low overlap between methods, recurring explanation patterns were still observed within individual sensitivity classes. Across multiple examples, Public documents consistently emphasised terms associated with external communication and distribution, while Internal documents repeatedly highlighted operational and organisational terminology. Similarly, Confidential and Strictly Confidential documents consistently exhibited strong focus on regulatory, governance and sensitivity-related terminology.

The recurring nature of these explanation patterns suggests that the model may learn relatively stable lexical associations for different sensitivity categories. The consistency of these patterns supports the plausibility of the model's reasoning process and indicates that the generated explanations are not entirely instance-specific.

At the same time, several inconsistencies were also observed. Certain structurally common tokens, repeated template elements and formatting-related terms occasionally received high importance scores despite having limited semantic relevance to the corresponding classification task. These observations suggest that parts of the learned representations may still be influenced by dataset-specific artifacts and repetitive structural patterns.

The controlled and synthetic nature of the dataset likely contributes to both the observed consistency and the identified limitations. Because the dataset contains relatively structured document templates and recurring vocabulary patterns, the model is encouraged to learn stable lexical associations across

classes. While this improves interpretability under experimental conditions, it may also reduce the diversity of explanation patterns compared to real-world financial documents.

Overall, the consistency analysis demonstrates that both explanation methods provide partially stable and semantically meaningful insights into the classification process. However, the observed divergence between methods also highlights the importance of interpreting explainability results with caution, particularly in regulated environments where explanation reliability and robustness are critical requirements.

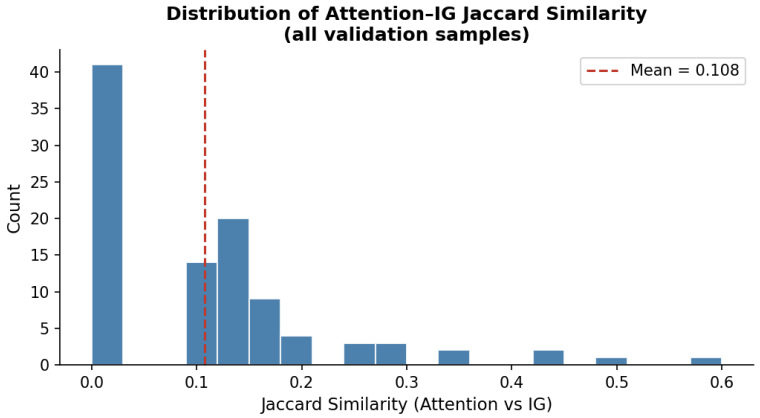

Figure 2: Distribution of Jaccard similarity scores between attention-based explanations and Integrated Gradients attributions across validation samples. The low average similarity indicates limited overlap between the explanation methods.

6.3 Confidence Analysis

In addition to evaluating predictive performance and explainability, the confidence of the model predictions was also analysed. Confidence analysis provides further insight into the certainty of the classification decisions and helps identify cases where the model may exhibit uncertainty or overconfidence.

The confidence scores were derived from the softmax probabilities produced by the classifier for each validation sample. Higher confidence values indicate that the model strongly favours a particular sensitivity class, whereas lower confidence values suggest increased ambiguity between competing classes.

Overall, the model exhibited relatively high confidence across a large portion of the validation samples. This behaviour is consistent with the high classification performance observed in previous sections and reflects the structured nature of

the synthetic dataset. In many correctly classified examples, the predicted class probability was strongly concentrated on a single sensitivity category, indicating that the model identified clear lexical and contextual signals associated with the corresponding class.

However, confidence analysis also revealed variations across different samples and sensitivity categories. Samples containing overlapping vocabulary or less explicit sensitivity indicators generally produced lower confidence scores compared to documents containing highly distinctive terminology. In particular, documents belonging to the Internal and Confidential categories occasionally exhibited increased uncertainty, reflecting the semantic similarity between these sensitivity levels.

Figure 3 illustrates the distribution of prediction confidence scores across the validation set. The distribution shows that a large proportion of predictions are associated with high confidence values, while a smaller subset of samples exhibits lower confidence levels. These lower-confidence predictions are primarily associated with documents containing mixed contextual signals or overlapping domain-specific terminology.

In addition to confidence scores, entropy-based uncertainty analysis was also performed. Entropy provides a measure of prediction uncertainty by considering the distribution of probabilities across all classes rather than only the highest predicted probability. Lower entropy values indicate highly confident predictions concentrated on a single class, whereas higher entropy values suggest increased uncertainty and ambiguity.

The entropy distribution demonstrates that most predictions are associated with relatively low uncertainty, further supporting the observation that the model frequently produces confident classification decisions under the controlled dataset conditions. At the same time, a smaller number of higher-entropy samples indicate cases where the model is less certain about the appropriate sensitivity category.

These observations suggest that the model confidence is strongly influenced by the presence of explicit lexical indicators and structured document patterns within the dataset. While high confidence values may indicate stable decision boundaries under controlled conditions, they do not necessarily guarantee robust generalization to more diverse real-world financial documents. In particular, high-confidence predictions may still result from reliance on superficial lexical cues rather than deeper semantic understanding.

Overall, the confidence analysis complements the explainability results by providing additional insight into the certainty and stability of the model's predictions. The observed relationship between confidence, lexical specificity and dataset structure further highlights the importance of carefully evaluating pre-

diction reliability in regulated financial environments.

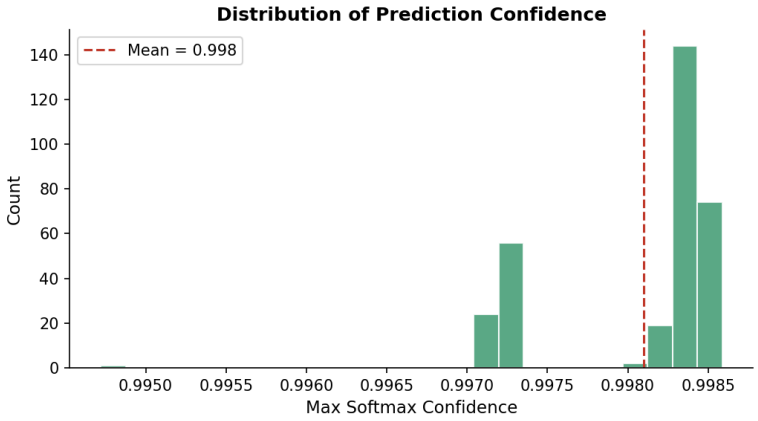

Figure 3: Distribution of prediction confidence scores across the validation samples. Most predictions exhibit high confidence values, indicating strong class certainty under the controlled dataset conditions.

6.4 Misclassification and Robustness Analysis

Although the classifier achieved perfect predictive performance on the validation set, additional analysis was conducted to assess the robustness and reliability of the learned classification behaviour. In particular, the analysis focuses on understanding the implications of the observed performance under the controlled experimental conditions used in this study.

The confusion matrix presented in Section 6.1.3 shows that the classifier correctly identified all evaluated samples across the four sensitivity categories. This indicates that the model successfully learned clear decision boundaries within the structured synthetic dataset and was able to consistently distinguish between the different sensitivity levels.

The absence of observed misclassifications suggests that the dataset contains relatively strong and distinguishable lexical and contextual signals associated with each class. In particular, the model appears to rely on recurring sensitivity-related terminology, organisational language and document structure patterns when producing classification decisions.

While these results demonstrate strong classification capability under controlled conditions, they do not necessarily guarantee robustness in more diverse real-world environments. The structured and synthetic nature of the dataset likely simplifies the classification task by reducing linguistic variability and increasing the consistency of class-specific patterns.

The explainability analysis presented in previous sections further supports this observation. Both attention-based explanations and Integrated Gradients attributions indicate that the model frequently focuses on explicit lexical indicators associated with individual sensitivity categories. This behaviour suggests that the model may partially rely on stable keyword associations and repetitive structural patterns when forming predictions.

As a result, the absence of misclassifications should be interpreted with caution. Perfect classification performance on synthetic validation data may reflect the controlled dataset design rather than complete semantic understanding of document sensitivity. In real-world financial environments, documents often contain ambiguous language, overlapping terminology, incomplete contextual information and more diverse writing styles, which may increase classification difficulty.

The robustness analysis therefore highlights an important limitation of the current experimental setup. Although the model demonstrates highly stable performance under controlled conditions, further evaluation using more diverse and less structured datasets would be necessary to assess generalization capability and real-world applicability.

Overall, the robustness evaluation suggests that the proposed approach is effective within the defined experimental setting while also emphasising the importance of careful interpretation of near-perfect classification performance in synthetic data environments.

While the above analysis provides insight into misclassification behaviour within the controlled validation dataset, it does not fully capture how the model performs under conditions that deviate from the training distribution. In practical deployment scenarios, documents will exhibit greater linguistic and structural variability than those represented in the controlled dataset used in this study. To further assess model robustness, additional experiments were conducted under conditions of increased variability, as described in the following subsection.

6.4.1 Robustness Testing Using External Documents

In addition to the misclassification patterns observed on the documents provided by the bank, further robustness analysis was conducted using six externally provided synthetic documents supplied by the collaborating bank. The same trained model and evaluation process was used without additional fine-tuning.

These documents differed substantially from the dataset used during training and validation. Although they remained synthetic and did not contain real confidential information, they exhibited greater linguistic and structural variability,

including less standardized formatting, more diverse sentence construction and less predictable contextual indicators of sensitivity. Unlike the controlled generation process used for the primary dataset, these documents were not constructed using the same templates, vocabulary constraints, or structural conventions.

Of the six documents, four were relatively similar in length and structure to the documents used in the primary dataset, consisting of approximately one-page documents, although they still exhibited greater variability. The remaining two documents differed substantially from the training data distribution, consisting of significantly longer and more structurally varied documents of approximately 50 pages each. As a result, the externally provided documents introduced a clear distributional shift relative to the training data and more closely resembled realistic organisational communication patterns.

When evaluated on these documents, the model demonstrated reduced robustness compared to the perfect performance observed on the original validation set. Among the four shorter documents, two *Strictly Confidential* documents were correctly classified, while one *Public* document was misclassified as *Confidential* and one *Internal* document was misclassified as *Strictly Confidential*. Of the two longer documents, one *Public* document was correctly classified, while one *Internal* document was misclassified as *Confidential*. These results indicate decreased generalization when exposed to documents with substantially different linguistic and structural characteristics, suggesting that part of the model's strong performance on the controlled dataset depends on regularities specific to the training data.

This degradation however does not necessarily imply limitations of the underlying transformer architecture. Rather, it highlights the importance of dataset diversity and robustness evaluation in regulated environments. While transformer-based models effectively learn contextual models, they may also become sensitive to recurring structures and lexical regularities present in controlled dataset. When these regularities change, model behaviour may become less stable.

These findings are consistent with earlier misclassification observations, reinforcing that the model relies in part on structural and lexical cues that are less prominent in more realistic and unconstrained document settings.

An important limitation of the evaluation is that the primary training and validation datasets were generated using the same controlled template-based document generation methodology. Although separate documents were used for training and validation, both datasets were derived from the same generation process and therefore share similar linguistic and structural characteristics. Consequently, the near-perfect performance observed on the primary evaluation dataset may partly reflect the model's ability to generalize within this controlled generation framework rather than to fully unconstrained financial documents. Further evaluation using independently collected datasets would therefore be

required to assess real-world generalisability.

6.5 Token Leakage / Shortcut Learning Analysis

The near-perfect classification performance observed throughout the evaluation raises important questions regarding the features learned by the model and the extent to which the predictions reflect genuine semantic understanding. To further investigate this behaviour, additional analysis was conducted to examine potential token leakage and shortcut learning effects within the dataset and model predictions.

Shortcut learning refers to situations where machine learning models rely heavily on superficial or easily identifiable patterns rather than developing deeper semantic representations of the underlying task. In text classification settings, this often occurs when specific keywords or repetitive lexical structures become strongly associated with individual classes during training.

The explainability results presented in previous sections suggest that the model frequently relies on explicit sensitivity-related terminology when producing classification decisions. Attention-based explanations and Integrated Gradients attributions repeatedly highlighted recurring class-specific tokens such as *public*, *salary*, *procurement*, *insider*, *disclosure* and *confidential*. These observations indicate that certain lexical cues may function as strong shortcuts for identifying sensitivity categories.

Figures 4 and 5 illustrate representative shortcut tokens associated with the Public and Strictly Confidential sensitivity categories. Public documents are strongly associated with externally oriented terminology such as *community*, *sustainability*, *investor*, *press* and *engagement*, whereas Strictly Confidential documents emphasise governance, strategic operations and regulatory terminology including *strategic*, *disclosure*, *divestiture*, *capital* and *strictly*.

The consistency of these token-level patterns across multiple explanation methods suggests that the model learns relatively stable lexical associations between specific terms and sensitivity classes. While these associations contribute to strong predictive performance within the controlled dataset environment, they also indicate the possibility that the model partially relies on shortcut features rather than broader contextual reasoning.

The structured and synthetic nature of the dataset likely amplifies this effect. Because many documents follow recurring templates and contain relatively predictable vocabulary distributions, certain tokens become highly informative for the classification task. As a result, the model may achieve high accuracy by identifying explicit lexical indicators instead of developing more generalized semantic representations of document sensitivity.

This behaviour highlights an important limitation of controlled synthetic datasets in explainability research. Although structured datasets improve interpretability and simplify analysis of model behaviour, they may also encourage reliance on shortcut features that do not generalize well to more diverse real-world documents.

At the same time, the observed shortcut patterns are not entirely unrealistic within the financial domain. In practice, sensitivity classifications are often associated with recurring regulatory, organisational and operational terminology. Therefore, some degree of lexical association is expected and may reflect meaningful domain characteristics rather than purely spurious correlations.

Overall, the token leakage and shortcut learning analysis demonstrates that the model's predictions are strongly influenced by recurring lexical cues and structured document patterns. These findings further emphasise the importance of combining predictive evaluation with explainability analysis when assessing AI-based document classification systems in regulated financial environments.

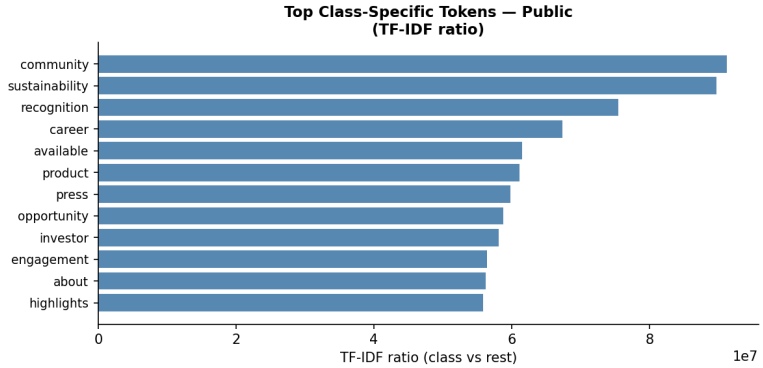

Figure 4: Class-specific shortcut tokens identified for the Public sensitivity category using TF-IDF ratio analysis. The highlighted terms are primarily associated with public communication, external engagement and publicly accessible organisational information.

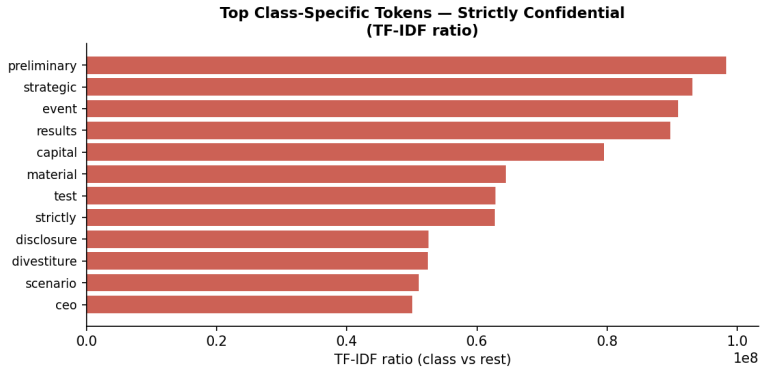

Figure 5: Class-specific shortcut tokens identified for the Strictly Confidential sensitivity category using TF-IDF ratio analysis. The highlighted terms are strongly associated with strategic, regulatory and governance-related content, indicating highly distinctive lexical patterns learned by the model.

7 Discussion

This chapter discusses the results presented in the previous chapter in relation to the research objectives and the broader challenges associated with AI-based document classification in regulated financial environments. In addition to evaluating predictive performance, the discussion focuses on the interpretability, consistency, robustness and practical applicability of the proposed transformer-based classification system integrated with explainability mechanisms.

Particular attention is given to the relationship between classification performance and explainability, including how attention-based explanations and Integrated Gradients attributions contribute to understanding the model's behaviour. The discussion also examines important limitations associated with synthetic and controlled data, shortcut learning, robustness and the potential challenges of deploying such systems in operational financial environments.

In addition, this chapter discusses the methodological limitations of the study, identifies areas for future research and summarizes the broader implications of the findings in relation to transparency, auditability and human oversight requirements within regulated financial domains such as banking and finance.

7.1 Interpretation of Classification Performance

The classification results show that transformer-based architectures can learn contextual and semantic patterns associated with information sensitivity levels in financial documents. The results provide evidence of strong predictive performance across the four compliance classifications, indicating that the contextual representations generated by the transformer architecture are suitable for distinguishing between the four sensitivity classes of documents.

The model appears to perform particularly well when processing documents containing clear contextual indicators of sensitivity. Many Strictly Confidential and Confidential documents contained language or references specific to the financial services industry, such as financial operations, internal strategies, personal information, or restricted organisational procedures. These contextual patterns appear to provide the transformer model with strong semantic associations related to higher sensitivity levels. This observation is consistent with prior studies showing that transformer-based architectures are effective at capturing contextual relationships and semantic dependencies within text [3, 6].

The findings also suggest that the model generalized beyond the presence of isolated sensitive terms. In other words, the classification behaviour indicates that the model learned broader contextual relationships across documents rather than relying solely on the existence of individual sensitive terms. Many sensitivity classifications appeared to depend on combinations of business terminology, document structure, intended audience and contextual framing. This is par-

ticularly relevant in regulated financial environments, where the sensitivity of information often depends on context and cannot easily be determined through rule-based keyword matching alone.

An important factor influencing classification performance was the structure of the dataset itself. Because the dataset was controlled and template-based, these characteristics likely contributed to more stable classification performance throughout the evaluation process. By reducing noise in the text and limiting uncontrolled variability within the data, the model was able to focus more directly on semantically meaningful patterns associated with information sensitivity. The controlled nature of the dataset also provided a useful setting for evaluating explainability methods and model behaviour under reproducible conditions. However, it may also have simplified aspects of the classification problem compared to fully unconstrained real-world financial data.

The robustness analysis suggests that the model was not heavily influenced by superficial surface-level features or isolated textual cues. Although some evidence of shortcut learning and token leakage was identified, the overall classification behaviour suggests that the model learned contextual relationships that were semantically meaningful across multiple document features.

This observation is consistent with the explainability analysis presented in Section 6.2. Both attention-based explanations and Integrated Gradients explanations generally highlighted semantically meaningful parts of the documents rather than arbitrary or unrelated tokens.

From a practical perspective, the results demonstrate that transformer-based models can provide a viable foundation for automating information sensitivity classification in regulated environments. The model successfully captured contextual representations relevant to organisational security policies while maintaining accurate predictions across multiple sensitivity levels.

Furthermore, the findings indicate that explainability mechanisms can be integrated into automated classification systems without causing substantial degradation in classification performance. This supports the broader objective of combining automation with transparency and auditability.

However, the findings must also be interpreted within the limitations of the study. Because the dataset consists of synthetic and controlled data, the reported performance should not be directly extrapolated to operational banking environments. Real-world financial documents contain significantly greater linguistic variability, ambiguity, inconsistent formatting and domain variation than represented in the experimental dataset. Consequently, although the results demonstrate the feasibility of the proposed approach, additional validation using real-world organisational data would be necessary before practical deployment could be considered.

7.2 Interpretation of Explainability Results

Section 6.2 provides insight into how the transformer-based classification model differentiates and characterizes varying levels of information sensitivity within financial documents. The integration of explainability mechanisms into the model not only improves transparency regarding the classification process, but also provides a means of evaluating whether predictions are based on semantically meaningful and contextually relevant information. Overall, the explainability analysis suggests that the model was generally able to produce classifications consistent with the predefined information sensitivity structure embedded within the dataset.

Both attention-based explanations and Integrated Gradients attributions identified tokens and phrases associated with indicators of financial sensitivity, organisational confidentiality and contextual references to restricted information. Across many of the analysed examples, the model consistently focused on concepts related to financial reporting, internal governance, strategic planning and confidential operational details. This indicates that the model relied not only on isolated keywords, but also on broader contextual relationships within documents when producing predictions.

The explainability analysis further demonstrates the importance of contextual interpretation in document classification. In many instances, highlighted tokens were not individually sensitive in isolation, but became important through their surrounding context. Generic financial terminology, for example, became increasingly significant when appearing alongside references to forecasts, internal resource planning, or acquisition discussions. This observation aligns with the theoretical foundations of transformer-based language models, where contextual relationships between tokens play a central role in representation learning and classification.

The attention-based explanations also provided a useful qualitative perspective on how the transformer architecture distributed focus across documents. In many examples, attention patterns were concentrated around semantically meaningful sections of text, particularly those involving confidentiality, strategic operations and financial metrics. This suggests that the self-attention mechanism was capable of capturing document-level relationships relevant to information sensitivity classification.

At the same time, the findings support prior research in the explainability literature indicating that attention cannot be treated as a fully faithful explanation of model reasoning. In several cases, attention weights appeared diffuse or distributed across tokens with limited semantic relevance to the classification outcome. Certain attention heads also emphasised structural or formatting-related tokens that were unlikely to contribute directly to the final prediction. These observations align with previous research showing that attention weights

do not necessarily correspond to causal importance within transformer models. Consequently, the attention visualisations in this study are better interpreted as indicators of model focus rather than definitive explanations of decision-making.

Integrated Gradients generally produced clearer and more concentrated attribution patterns compared to attention-based explanations. The attribution results frequently highlighted tokenized representations associated with security-relevant concepts such as confidential projects, internal budget allocations, customer records and restricted communications. Unlike attention weights, which often reflected broader contextual relationships across tokens, Integrated Gradients more directly identified which textual components contributed most strongly to the classification outcome.

The token-level attributions generated using Integrated Gradients further suggest that the model relied on combinations of linguistic cues rather than isolated trigger words. Terms such as "forecast," "merger," or "internal review" typically received stronger attribution scores when appearing alongside semantically related information concerning financial operations or organisational secrecy. This may indicate that the model learned contextual sensitivity patterns rather than simply memorizing keyword associations for each classification category.

One significant observation is that the generated explanations appeared broadly consistent with human expectations regarding information sensitivity. Documents classified as Public generally emphasised externally oriented language, recruitment-related material, or publicly available reporting. Internal documents focused more heavily on operational and governance-related terminology, while Confidential and Strictly Confidential documents emphasised customer-related data, financial projections, strategic planning and proprietary business information. This alignment between model explanations and expected organisational sensitivity indicators increases confidence that the model learned semantically meaningful distinctions between sensitivity classes.

At the same time, the explainability analysis revealed several limitations and ambiguities. Overlapping explanation patterns were frequently observed between neighboring sensitivity categories, particularly between Internal and Confidential documents. Since many operational terms appeared across multiple sensitivity levels, attribution patterns occasionally became less distinct. This suggests that the model's internal representation of sensitivity levels is not entirely discrete, but instead reflects gradual contextual differences between classes. Such overlap is realistic in practical document classification settings, where sensitivity boundaries are often ambiguous even for human reviewers.

The relatively high consistency and clarity observed in the explanations were likely influenced by the controlled and synthetic nature of the dataset. Since the dataset intentionally constrained linguistic variability and document structure, the model was able to learn more stable relationships between textual

features and sensitivity labels. As a result, attribution patterns became easier to interpret and less affected by irrelevant linguistic noise. While this improves methodological clarity, it also means that the explanations may appear cleaner and more stable than would likely be the case in fully unconstrained real-world financial documents.

Another important finding concerns the role of explainability in supporting transparency rather than proving correctness. The generated explanations do not guarantee that the model reasons in the same manner as a human expert, nor do they provide formal proof that the model is free from latent biases or shortcut learning. Instead, the explanations function primarily as diagnostic and inspection tools that allow human reviewers to evaluate whether the model appears to rely on plausible and policy-relevant information. In regulated financial environments, this form of transparency is valuable because it supports auditability, manual review and organisational oversight.

Overall, the explainability results demonstrate that post-hoc explanation methods can provide meaningful insight into model behaviour without requiring modification of the underlying transformer architecture. The combination of attention-based explanations and Integrated Gradients attributions provided complementary perspectives on the model's reasoning, enabling both document-level and token-level inspection of classification decisions. Although these explanations do not eliminate the opacity of deep learning systems, the findings suggest that integrating explainability mechanisms into AI-assisted document classification systems can substantially improve interpretability and practical trustworthiness within regulated financial environments.

7.3 Consistency Between Explanation Methods

The primary goal of this study was not only to produce explanations for the model's classification decisions, but also to determine whether different explainability methods focused on the same contextual areas of a document when generating explanations for a specific classification decision. Therefore, explanations generated through attention-based indicators and Integrated Gradients feature attributions were analysed against one another.

The analyses showed that both methods frequently identified the same contextual areas within documents, especially when the model produced classifications with high confidence. In addition, both methods frequently highlighted domain-specific terminology, references to confidentiality, financial terminology and organisational indicators that contributed to determining the sensitivity level of a document. For instance, terms such as "Prediction Ranges" related to internal financial forecasts, "Operational Details" referring to non-public operational information, client-related terminology and explicit confidentiality markers were often assigned similar importance across both explanation methods. This suggests that the model based its classifications on semantically meaningful textual

signals rather than on random or unrelated tokens.

The comparison between the two explanation methods also revealed important differences. Consistent with the theoretical understanding of the two approaches, attention-based explanations generally reflected broader contextual relationships between sentences and different parts of the document, while Integrated Gradients produced more localized token-level attributions focused on individual words or phrases. These differences suggest that the two methods provide complementary perspectives on how sensitive information is classified by the transformer model. Integrated Gradients approximates the contribution of individual tokens to the final prediction, whereas attention weights reflect how the model distributes focus across contextual relationships within the document. Consequently, agreement between the two methods can be interpreted as an indicator of explanation reliability, while their differences demonstrate that the methods provide complementary rather than identical explanations. This finding supports the conclusion that the transformer model learned both local lexical patterns and broader contextual patterns related to identifying sensitive information.

However, consistency between the explanation methods was not uniform across all documents. While the controlled validation dataset generally produced highly confident and correct classifications, weaker overlap between attention-based explanations and Integrated Gradients attributions was more commonly observed during robustness testing and in the limited number misclassified examples. In many such cases, the generated explanations emphasised more generic organisational language or structurally prominent tokens rather than tokens directly associated with sensitive information. This behaviour may reflect uncertainty in the model's classification process, as well as a greater reliance on weaker or more ambiguous contextual signals. Prior research suggests that alignment between explanation methods may become less consistent when models operate near decision boundaries or lack strong evidence for classification. The limited inconsistencies observed in this study were primarily associated with external robustness-testing examples and a small number of challenging classification cases.

In general, the controlled and structured design of the dataset likely contributed to the strong consistency observed across many documents. The consistency between explanation methods can be attributed to the constrained vocabulary, structure and contextual patterns used during dataset construction. As a result, the model was able to learn more stable relationships between textual cues and sensitivity labels. In addition, the reduced variability across documents may have lowered the level of noise in the generated explanations, making the explanations easier to interpret and compare across methods. At the same time, these findings raise questions regarding whether similarly strong consistency between explanation methods would also be observed in real-world financial documents containing greater linguistic diversity and ambiguity.

From a regulatory and organisational perspective, consistency between explanation methods is important because it strengthens confidence in the transparency of the system. In regulated financial environments, relying on a single explanation mechanism may be insufficient for establishing trust or supporting auditability. By employing complementary explanation methods, multiple perspectives on the model's decision-making process can be provided, reducing the risk that explanations are artifacts of a single explanation mechanism. When both explanation methods consistently highlight semantically meaningful information, human reviewers may have greater confidence that the classification decision is grounded in relevant contextual evidence.

However, consistency alone should not be interpreted as proof that a model is fully faithful or correct. Previous studies have shown that different explanation methods may agree even when a model partially relies on spurious correlations or dataset-specific shortcuts. Consequently, consistency should be treated as one indicator of reliability rather than as definitive validation of the model's reasoning process. For this reason, consistency is interpreted together with the robustness testing, misclassification analysis and shortcut learning investigations presented in later sections.

Overall, the findings demonstrate that combining attention-based explanations with Integrated Gradients provides a more comprehensive understanding of model behaviour than relying on either method individually. The overlap between the two methods supports the conclusion that the classifier generally bases its predictions on meaningful contextual indicators related to document sensitivity, while the differences between the methods provide additional insight into both contextual dependencies and token-level feature importance.

7.4 Shortcut Learning, Token Leakage and Robustness Considerations

When assessing the outputs of the proposed document classification system, an important consideration is whether shortcut learning and/or token leakage occurred within the dataset. Although the model demonstrated strong classification performance and often produced explanations that appeared consistent with the semantic content of the documents, these results do not necessarily guarantee that the model developed robust contextual representations of information sensitivity. Transformer-based language models are highly effective at identifying statistical regularities within data, including superficial patterns and correlations that may not generalize beyond the training environment.

Shortcut learning occurs when a machine learning model relies on simple predictive characteristics or highly correlated features to minimize training loss, rather than learning deeper semantic relationships relevant to the intended task. In the context of document classification, shortcut learning may involve exces-

sive reliance on specific keywords, repeated phrases, formatting conventions, or structural patterns that strongly correlate with particular document classes. Examples of such indicators may include confidentiality notices, internal communication markers, or recurring financial terminology. If these features appear disproportionately within certain classes of documents, the model may assign excessive importance to them during classification.

Since the dataset used in this study is synthetic and controlled, shortcut learning represents a particularly important risk factor. The dataset generation process intentionally constrained linguistic variability and document structure in order to support controlled experimentation and improve explainability analysis. While this design provided methodological benefits for the overall goals of the study, it may also have unintentionally strengthened correlations between specific tokens and sensitivity labels. As a result, the model may have partially relied on recurring lexical indicators rather than learning broader contextual or semantic relationships between different document elements.

The explainability analysis conducted as part of this study provided some insight into this challenge. In particular, both attention-based explanations and Integrated Gradients feature attributions frequently highlighted tokens and phrases that were intuitively associated with document sensitivity, including references to financial information, access restrictions, internal processes and confidentiality-related terminology. Many of these highlighted elements appeared semantically meaningful and aligned with what would reasonably be expected in the context of information classification within financial documents. However, the presence of plausible explanations does not necessarily indicate that the underlying reasoning process is robust or causally grounded in genuine contextual understanding.

Additionally, the explainability methods revealed indications that some predictions may have been influenced by localized lexical cues or repeated structural patterns. In several cases, highly weighted tokens corresponded to terms strongly associated with particular document classes and may therefore have functioned as shortcut indicators within the dataset. This highlights an important limitation of explainability methods: while they can help identify potentially problematic model behaviour, they cannot by themselves confirm that the model has developed generalized reasoning capabilities. Explainability methods should therefore be viewed primarily as diagnostic tools that support the inspection and evaluation of model behaviour, rather than as definitive evidence of reliable reasoning.

The robustness experiments performed using external texts partially reduce these concerns, although the results also revealed limitations in the model's ability to generalize beyond the controlled training environment. While the model was able to correctly classify documents that were not explicitly represented in the training dataset, multiple misclassifications also occurred, indicating that

the learned representations were not fully robust. These findings suggest that the classification decisions were not based solely on memorization of template-specific patterns, but they also demonstrate that the model may still rely on dataset-specific correlations that do not generalize consistently to external document flows. Additionally, attention-based explanations and Integrated Gradients attributions were generally consistent across multiple documents, indicating that the model may have learned at least partially stable sensitivity-related signals rather than relying entirely on random artifacts. Nevertheless, it remains possible that hidden dataset-specific correlations or residual shortcut behaviours still exist and the observed results should therefore not be interpreted as conclusive evidence of robust reasoning.

From a practical perspective, shortcut learning represents a significant concern when considering the deployment of AI-based classification systems within regulated financial environments. A model that primarily relies on surface-level lexical cues may achieve high performance on controlled datasets while failing to generalize effectively to operational document flows characterized by greater linguistic diversity, ambiguity, or adversarial variation. Such behaviour could result in incorrect classifications, reduced trustworthiness and challenges during regulatory review or audit processes. In high-stakes domains such as finance and banking, these risks highlight the importance of continuous validation, human oversight and ongoing monitoring of model behaviour after deployment.

7.5 Controlled Data and Methodological Limitations

The use of synthetic and controlled data has important implications for both the strengths and limitations of the proposed document classification system. Within regulated financial environments, direct access to real-world documents is highly restricted due to privacy, confidentiality and regulatory constraints. The use of synthetic data therefore enabled experimentation with transformer-based models and explainability techniques while avoiding exposure of sensitive information and ensuring compliance with organisational and legal requirements.

An important advantage of the controlled dataset design is that it reduces linguistic variability and noise, allowing the model to focus more directly on contextual and semantic patterns associated with information sensitivity classification. This controlled setting also supports explainability analysis, as clearer relationships can be observed between textual features and model predictions. As discussed throughout the explainability analysis, the reduced variability of the dataset contributed to more interpretable attribution patterns and more consistent explanation behaviour across similar documents.

At the same time, the controlled nature of the dataset introduces important methodological limitations. Real-world financial documents often contain significantly greater linguistic diversity, ambiguity, inconsistent formatting and

domain-specific complexity than represented in the synthetic dataset used in this study. As a result, the classification performance and explanation consistency observed in this work may not fully generalize to operational financial environments.

The template-based generation process may also introduce structural regularities that are easier for the model to learn than naturally occurring language patterns. Although robustness testing and shortcut learning analysis were conducted to investigate this issue, the possibility remains that some classification decisions are influenced by artifacts of the controlled generation process rather than by fully generalizable semantic understanding. This limitation is particularly important when interpreting the high classification performance achieved by the model.

In addition, the study is limited to English-language internal financial documents and does not evaluate multilingual settings or broader document categories commonly encountered in financial institutions. The explainability analysis is also restricted to attention-based explanations and Integrated Gradients feature attribution. While these techniques provide complementary perspectives on model behaviour, they do not represent the full range of available explainability approaches and alternative methods may produce different insights regarding model interpretability and faithfulness.

Another limitation concerns the absence of human-centered evaluation. Although the generated explanations were qualitatively assessed in relation to plausibility, consistency and semantic relevance, the study does not include formal user studies involving compliance officers, auditors, or information security specialists. Consequently, the work does not directly measure user trust, explanation usability, or organisational acceptance in real-world operational contexts.

Furthermore, the proposed system is evaluated within a controlled experimental setting rather than within a production environment. Factors such as integration with organisational workflows, scalability, long-term model maintenance and evolving regulatory requirements are therefore outside the scope of this thesis.

Despite these limitations, the controlled experimental design provides important methodological advantages. It enables systematic evaluation of transformer-based document classification and explainability techniques under conditions where privacy-preserving experimentation would otherwise be difficult to achieve. The study therefore demonstrates the feasibility of combining explainable AI methods with automated document classification while highlighting the challenges that remain for deployment in real-world regulated financial environments.

7.6 Applicability in Regulated Financial Environments

The findings of this study suggest that transformer-based document classification systems combined with explainability techniques may provide practical value within regulated financial environments. In particular, the results indicate that contextual language models are capable of identifying patterns associated with information sensitivity classification while simultaneously providing explanations that support transparency and human inspection of model behaviour.

In regulated financial organisations, explainability is particularly important because automated decisions often require justification to internal stakeholders, auditors and regulatory authorities. The use of attention-based explanations and Integrated Gradients feature attribution demonstrates how explainability mechanisms can provide insight into which parts of a document contribute most strongly to classification outcomes. Such explanations may support compliance processes, manual review procedures and audit-related activities by allowing human reviewers to inspect and validate automated decisions.

At the same time, the results suggest that explainable AI systems should primarily be viewed as decision-support tools rather than fully autonomous classification systems. Although the model achieved strong predictive performance within the controlled experimental setting, explainability remains essential for enabling human oversight and identifying potentially unreliable or misleading model behaviour. This is particularly important in high-stakes environments where incorrect classification of sensitive information may result in legal, financial, or reputational consequences.

The observed consistency between explanation methods also indicates potential value for transparency-oriented workflows. When attention-based indicators and feature attribution methods highlight similar contextual patterns, reviewers may gain increased confidence that the classification model is relying on semantically meaningful information rather than on irrelevant correlations or shortcut features.

From an organisational perspective, the study demonstrates that explainability mechanisms can be integrated into transformer-based document classification systems without fundamentally changing the underlying model architecture. This is relevant for regulated financial environments, where practical deployment often depends not only on predictive performance, but also on reproducibility, auditability and compatibility with existing governance processes.

However, practical deployment would still require additional validation using more diverse and operationally realistic datasets, together with human-centered evaluation involving relevant domain experts. Consequently, the results of this thesis should primarily be interpreted as demonstrating the feasibility and potential applicability of explainable AI-based document classification in regulated

financial settings rather than as evidence of production readiness.

7.7 Future Work

Several directions for future research emerge from the findings and limitations of this study. One of the most important areas for future work involves evaluating the proposed approach using more diverse and operationally realistic financial documents. While the use of synthetic and controlled data enabled controlled experimentation under regulatory constraints, additional evaluation using real-world enterprise documents would provide stronger insight into the practical robustness and generalizability of the proposed system.

Future research could also investigate multilingual document classification scenarios. Financial institutions frequently operate across international environments where documents are produced in multiple languages and formats. Extending the proposed framework to multilingual settings would provide a more realistic representation of operational financial workflows and introduce additional challenges related to linguistic diversity and contextual variation.

Future work could also explore more advanced dataset generation methodologies designed to balance linguistic diversity with experimental control. One possible direction involves the use of Grammatical Framework (GF) [28] based text generation approaches to produce datasets containing more varied and complex sentence structures while still maintaining controllable semantic and contextual properties. Such approaches may support more rigorous robustness evaluation and allow explainability analyses to be conducted under more realistic linguistic conditions.

In addition, future research may involve closer collaboration with financial institutions in order to evaluate the proposed approach using operational organisational data. This may include the use of properly anonymized enterprise documents or alternatively enabling model training directly within secure internal organisational environments where sensitive data cannot leave the institution. Such approaches could improve both realism and practical applicability while remaining compliant with regulatory and organisational security requirements.

Another important direction concerns the expansion of explainability techniques beyond attention-based explanations and Integrated Gradients. Future studies could compare additional explainability approaches in order to better understand the strengths and limitations of different explanation mechanisms within regulated NLP applications.

The robustness evaluation could also be extended through larger-scale adversarial and out-of-distribution testing. Future experiments may investigate how

transformer-based classification systems respond to noisy input data, ambiguous language, manipulated terminology, formatting variations, or intentionally adversarial document modifications designed to challenge sensitivity classification systems.

Further work may additionally explore human-centered evaluation of explainability. While this study primarily focused on technical explainability analysis, future research could investigate how compliance officers, auditors and information security specialists interpret and use generated explanations during practical document review processes. Such evaluations would provide valuable insight into the usability and organisational trustworthiness of explainable AI systems within regulated environments.

From a practical perspective, future work may also investigate deployment-oriented considerations such as scalability, computational efficiency, model monitoring, human-in-the-loop workflows and continuous adaptation to evolving organisational terminology and regulatory requirements.

Finally, future research could investigate hybrid approaches combining contextual transformer models with rule-based or policy-driven classification mechanisms. Such approaches may help balance predictive flexibility with regulatory transparency and improve robustness in operational financial environments.

Overall, future work should continue exploring how explainable AI techniques can support trustworthy, transparent and practically applicable document classification systems within regulated financial domains.

7.8 Summary

This chapter discussed the findings of the study in relation to explainability, robustness, transparency and the practical applicability of AI-based document classification within regulated financial environments. The discussion highlighted that transformer-based classification models combined with explainability mechanisms can provide meaningful insight into document sensitivity classification decisions while also improving interpretability and transparency.

The results demonstrated that both attention-based explanations and Integrated Gradients attributions were capable of identifying semantically meaningful contextual indicators associated with different sensitivity levels. At the same time, the analyses revealed important limitations related to shortcut learning, dataset structure, robustness and the use of synthetic and controlled data.

The discussion further emphasised that explainability methods should not be interpreted as definitive proof of faithful model reasoning, but rather as diagnostic tools that support inspection and evaluation of model behaviour. In addition, the findings highlighted the importance of combining multiple expla-

nation methods, robustness analysis and human oversight when evaluating AI systems intended for regulated financial environments.

Several methodological limitations were also identified throughout the discussion. In particular, the inability to obtain anonymized operational financial documents resulted in the use of synthetic and controlled datasets, which likely influenced both classification performance and explanation consistency. Although the controlled experimental setup enabled systematic evaluation of explainability techniques, the observed results may not fully generalize to operational enterprise environments characterized by greater linguistic and structural variability.

The discussion additionally explored the practical applicability of explainable transformer-based document classification systems within regulated industries. The findings suggest that such systems may provide valuable decision-support capabilities for information classification workflows while supporting transparency, auditability and human oversight requirements.

Overall, the study demonstrates that explainable transformer-based document classification systems have promising potential to support transparency, auditability and decision-support workflows within regulated financial environments. At the same time, the findings reinforce the importance of careful evaluation of robustness, interpretability, shortcut learning behaviour and generalization when applying AI systems to high-stakes financial and regulatory contexts.

8 Conclusion

This thesis investigated the use of transformer-based language models combined with explainable artificial intelligence techniques for automatic document classification in regulated financial environments. The study was motivated by the increasing volume of digital information handled within financial institutions and the associated challenges related to information security, regulatory compliance and efficient information management. In such environments, incorrect handling or classification of sensitive information may result in legal, financial and reputational consequences. While transformer-based models have demonstrated strong performance in document classification tasks, their adoption in regulated domains remains challenging due to concerns related to transparency, auditability and the black-box nature of modern machine learning systems.

The primary objective of this thesis was to investigate whether transformer-based language models can effectively classify documents according to information sensitivity levels while also providing explanations that support transparency and human oversight. In particular, the study examined how explainability mechanisms can contribute to transparency, trust and auditability in AI-based document classification systems operating within regulated financial contexts. To address these objectives, a transformer-based classification model was developed and evaluated using a synthetic and controlled dataset representing different levels of information sensitivity. Explainability was integrated into the system through the use of attention-based indicators and Integrated Gradients feature attribution methods.

The results of the study demonstrate that transformer-based language models are capable of learning contextual and semantic patterns associated with information sensitivity classification. The model successfully distinguished between different security levels, including Public, Internal, Confidential and Strictly Confidential documents. The findings further suggest that the model relied not only on isolated keywords, but also on contextual relationships and structural characteristics within documents when performing classification. This supports the suitability of transformer-based architectures for document-level classification tasks in regulated environments where contextual interpretation is essential.

The explainability analysis demonstrated that both attention-based explanations and Integrated Gradients attributions were able to highlight linguistically and semantically meaningful parts of the input documents. The explanations frequently corresponded to contextual indicators associated with sensitive information, including financial terminology, references to internal operations and regulatory or confidential content. In addition, observations of explanation consistency across documents with similar characteristics indicated relatively stable model behaviour. The use of multiple complementary explanation techniques also provided a broader perspective on the model's decision-making process and reduced reliance on a single interpretability method.

At the same time, the study identified several limitations. The controlled and synthetic nature of the dataset, while necessary due to privacy and regulatory constraints, reduced linguistic variability and may have contributed to shortcut learning and reliance on dataset-specific patterns. Furthermore, both the training and evaluation data originated from the same generation process, limiting the ability to assess generalization to operational financial environments. Consequently, the results should be interpreted within the context of the controlled experimental setting, and further evaluation on independent and more diverse datasets is required to assess the robustness and practical applicability of the proposed approach.

Despite these limitations, the study demonstrates the practical feasibility of combining transformer-based document classification with explainability techniques in regulated financial settings. The proposed approach shows potential as a decision-support tool capable of assisting information security specialists, compliance officers and other stakeholders in document handling workflows. Rather than replacing human oversight, the system is intended to support transparency, inspection and informed decision-making in environments where accountability and auditability are critical requirements.

Future work may extend this research by evaluating the proposed approach on more diverse and realistic financial datasets, including multilingual and less controlled document collections. Additional explainability methods and human-centered evaluation approaches may also provide deeper insight into how explanations are interpreted and used by domain experts in practice. Furthermore, future studies may investigate methods for improving robustness against shortcut learning and for evaluating explanation faithfulness more systematically.

In conclusion, this thesis demonstrates that explainable transformer-based language models can support automatic information sensitivity classification in regulated financial environments while also providing meaningful transparency mechanisms for human inspection and auditability. The findings suggest that explainability should not be treated as an optional addition to AI-based systems operating in regulated domains, but rather as a core requirement for enabling trustworthy and accountable automated decision-making.

9 Ethics

There are several ethical considerations in this research, particularly in relation to data privacy, the use of pretrained artificial intelligence models and the use of automated decision-support systems in regulated financial environments.

Firstly, in relation to the use of AI models in the production of this thesis, ChatGPT was used to refine language and improve expression. All generated content was reviewed, validated and edited to ensure academic integrity, correctness and appropriateness in relation to the research objectives and goals.

The research employs a dataset comprising synthetic documents and samples generated in collaboration with a Swedish bank. In view of the sensitive nature of financial documents, direct access to internal data is not feasible. Consequently, synthetic data is generated to mimic the structural and linguistic features of financial documents without including any identifiable or confidential content. This approach is taken to ensure compliance with applicable data protection regulations, including the General Data Protection Regulation (GDPR) [1].

The language model applied in the project is based on the transformer architecture and is pretrained on publicly available data. Although fine-tuning of the model is applied only to synthetic data, it is important to note that biases present in the original training data may still influence the model's behaviour. Although bias mitigation is not the main focus of the thesis, this limitation is acknowledged during the evaluation process.

Furthermore, it is recognised that automated document classification may have significant implications for organisations and society. Incorrect classification of sensitive information may result in regulatory non-compliance, information leakage, or unnecessary restriction of information. For this reason, the proposed system is intended to function as a decision-support tool, with human oversight remaining an essential component of its practical application.

Finally, it is important to acknowledge that the explainability mechanisms applied in the system may also present limitations. Explanation techniques, including attention-based indicators and feature attribution methods, may lead to over-interpretation or may not fully reflect the model's internal reasoning. These limitations are therefore considered when analysing and presenting explanation results.

10 Acknowledgments

We would like to express our sincere gratitude to our supervisor for the valuable guidance and support during the project. In addition, we would also like to thank our industry contact and the collaborating banking institution for providing domain expertise and facilitating access to relevant resources under appropriate confidentiality constraints.

References

[1] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance). http://data.europa.eu/eli/reg/2016/679/oj, May 2016.

[2] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA relevance). http://data.europa.eu/eli/reg/2024/1689/oj, Jul 2024.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT*, 2019.

[4] Dogu Araci. FinBERT: Financial Sentiment Analysis with Pre-trained Language Models. In *Proceedings of the 1st Financial Narrative Processing Workshop (FNP 2019)*, 2019. doi: 10.18653/v1/W19-5506.

[5] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI. *Information Fusion*, 2020. doi: 10.1016/j.inffus.2019.12.012.

[6] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017. URL `https://arxiv.org/abs/1703.01365`.

[7] U.S. Congress / Federal Deposit Insurance Corporation. Financial Services Modernization Act of 1999 (Gramm–Leach–Bliley Act). `https://www.fdic.gov/consumer-compliance-examination-manual/viii-1-gramm-leach-bliley-act-privacy-consumer-financial`, 1999. Privacy of Consumer Financial Information; FDIC Consumer Compliance Manual.

[8] European Parliament and Council of the European Union. Directive 2014/65/EU of the European Parliament and of the Council on markets in financial instruments (MiFID II). `https://eur-lex.europa.eu/eli/dir/2014/65/oj`, 2014. Official Journal of the European Union L 173, 12 June 2014.

[9] Sandra Wachter, Brent Mittelstadt, and Luciano Floridi. Why a right to explanation of automated decision-making does not exist in the general data protection regulation. *International Data Privacy Law*, 7(2):76–99, 2017. doi: 10.1093/idpl/ipx005.

[10] Blerim Fazlija, Muhamed Ibraimi, Amir Forouzandeh, and Arta Fazlija. Reasoning with financial regulatory texts via large language models. *Journal of Behavioral and Experimental Finance*, 2025. doi: 10.1016/j.jbef.2025.101067.

[11] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019. doi: 10.1038/s42256-019-0048-x.

[12] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM Computing Surveys*, 2018. doi: 10.1145/3236009.

[13] European Commission High-Level Expert Group on Artificial Intelligence. Ethics guidelines for trustworthy ai, 2019. URL `https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai`. Accessed: 2026-05-21.

[14] Y. Bengio, P. Simard, and P. Frasconi. Learning long-term dependencies with gradient descent. *IEEE Transactions on Neural Networks*, 1994. doi: 10.1109/72.279181.

[15] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 1997. doi: 10.1162/neco.1997.9.8.1735.

[16] Y. Kim. Convolutional neural networks for sentence classification. In *EMNLP*, 2014.

[17] Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, and Ion Androutsopoulos. Large-Scale Multi-Label Text Classification on EU Legislation. In *EMNLP-IJCNLP*, 2019.

[18] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. Hierarchical attention networks for document classification. In *NAACL-HLT*, 2016.

[19] A. Vaswani et al. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.

[20] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017. doi: 10.48550/arXiv.1702.08608.

[21] Kunal Mishra, Hrishikesh Pagare, and Kunal Sharma. A Hybrid Rule-Based NLP and Machine Learning Approach for PII Detection and Anonymization in Financial Documents. *Scientific Reports*, 2025. doi: 10.1038/s41598-025-04971-9.

[22] Jayakumar Muralitharan and Chitra Arumugam. Privacy BERT-LSTM: A Novel NLP Algorithm for Sensitive Information Detection in Textual Documents. *Neural Computing and Applications*, 2024. doi: 10.1007/s00521-024-09707-w.

[23] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016. doi: 10.1145/2939672.2939778.

[24] Christoph Molnar. Interpretable machine learning: A guide for making black box models explainable. *arXiv preprint arXiv:2203.13788*, 2022.

[25] Sarthak Jain and Byron C. Wallace. Attention is not explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, 2019. doi: 10.18653/v1/N19-1357.

[26] Niklas Bussmann, Paolo Giudici, Davide Marinelli, and Jochen Papenbrock. Explainable ai in fintech risk management. *Frontiers in Artificial Intelligence*, 2021. doi: 10.3389/frai.2021.659619.

[27] Tobias Kuhn. A survey and classification of controlled natural languages. *Computational Linguistics*, 40(1):121–170, 03 2014. ISSN 0891-2017. doi: 10.1162/COLI_a_00168. URL `https://doi.org/10.1162/COLI_a_00168`.

[28] Aarne Ranta. Grammatical framework: A type-theoretical grammar formalism. *Journal of Functional Programming*, 14(2):145–189, 2004. doi: 10.1017/S0956796803004738.

[29] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations*, 2015. doi: 10.48550/arXiv.1412.6980.

[30] Helena Löfström, Karl Hammar, and Ulf Johansson. A meta survey of quality evaluation criteria in explanation methods. *arXiv preprint arXiv:2203.13929*, 2022. URL `https://arxiv.org/abs/2203.13929`.

[31] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

[32] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[33] Jianlong Zhou, Amir H Gandomi, Fang Chen, and Andreas Holzinger. Evaluating the quality of machine learning explanations: A survey on methods and metrics. *Electronics*, 2021. doi: 10.3390/electronics10050593.

Appendix

Controlled Template-Based Dataset Generation

Due to confidentiality, privacy and regulatory constraints, real financial documents could not be used for training and evaluation. Instead, a synthetic dataset was constructed using a controlled template-based document generation process. The approach imposed constraints on document structure, vocabulary, and contextual patterns in order to reduce variability and support systematic experimentation. Although the approach shares some characteristics with controlled language methodologies, it does not constitute a formal CNL and does not rely on a formal grammar or controlled language framework.

The dataset generation process was designed to reduce linguistic variability while preserving contextual and semantic characteristics relevant to information sensitivity classification. Documents were generated using predefined templates corresponding to the four information sensitivity levels used in this study: Public, Internal, Confidential and Strictly Confidential.

The generation process employed controlled vocabulary, predefined document structures and class-specific contextual indicators. Templates were populated using context-appropriate terminology and randomized placeholders representing entities such as departments, projects, customer information, financial activities and internal business operations. This approach enabled the creation of large numbers of synthetic documents while avoiding the use of real confidential information.

Unlike formal CNL systems, the approach used in this study does not enforce a strict grammar or formal semantic representation. Instead, it uses lightweight template-based generation with constrained linguistic variation. The objective was to create a controlled experimental environment suitable for explainability analysis while maintaining sufficient realism for information sensitivity classification.

Because both the training and evaluation datasets were generated using the same controlled generation methodology, they share similar linguistic and structural characteristics. Consequently, the results reported in this thesis should be interpreted within the context of this controlled experimental setting. Future work should evaluate the proposed approach using independently collected datasets and more diverse document structures in order to assess generalizability in real-world financial environments.

Figure 10 shows an example document generated using the synthetic dataset generation process.

Images

```
===== CORRECT PUBLIC =====

Document:
 Operational efficiency remains a key focu

=== TOP ATTENTION TOKENS ===
transparency    0.1497
document        0.1451
public          0.1298
[SEP]           0.1238
supports        0.1122
the             0.0927
[CLS]           0.0915
initiatives     0.0871
.               0.0489
management      0.0030

=== TOP IG TOKENS ===
transparency    1.7100
supports        1.5258
public          1.4964
initiatives     1.0146
document        0.6341
remains         0.3890
remains         0.3331
the             0.3019
[SEP]           -0.2684
.               0.2560
```

Figure 6: Analysis output for one random public document

```
===== CORRECT STRICT CONFIDENTIAL =====

Document:
 The quarterly financial report summarizes

=== TOP ATTENTION TOKENS ===
unreleased      0.1715
.               0.1343
.               0.1208
[CLS]           0.1075
figures         0.1057
earnings        0.0997
are             0.0992
included        0.0986
[SEP]           0.0505
departments     0.0082

=== TOP IG TOKENS ===
unreleased      2.2729
[SEP]           -1.7949
figures         1.4909
earnings        0.7706
performance     0.5051
.               -0.3274
continue        0.3207
sum             0.3166
the             0.2681
operational     0.2613
```

Figure 7: Analysis output for one random strictly confidential document

```
===== CORRECT PUBLIC =====

Document:
 Operational efficiency remains a k

=== TOP ATTENTION TOKENS ===
transparency    0.1497
document        0.1451
public          0.1298
[SEP]           0.1238
supports        0.1122
the             0.0927
[CLS]           0.0915
initiatives     0.0871
.               0.0489
management      0.0030

=== TOP IG TOKENS ===
transparency    1.7100
supports        1.5258
public          1.4964
initiatives     1.0146
document        0.6341
remains         0.3890
remains         0.3331
...
supports 1.5257724523544312
public 1.4963523149490356
initiatives 1.0145753622055054
document 0.6340627670288086
```

Figure 8: Analysis output for the public document in Figure 6 after masking the most influential tokens

```
===== CORRECT STRICT CONFIDENTIAL =====

Document:
 The quarterly financial report summarizes

=== TOP ATTENTION TOKENS ===
unreleased      0.1715
.               0.1343
.               0.1208
[CLS]           0.1075
figures         0.1057
earnings        0.0997
are             0.0992
included        0.0986
[SEP]           0.0505
departments     0.0082

=== TOP IG TOKENS ===
unreleased      2.2729
[SEP]           -1.7949
figures         1.4909
earnings        0.7706
performance     0.5051
.               -0.3274
continue        0.3207
...
figures 1.490944743156433
earnings 0.7706027626991272
performance 0.5051019191741943
. -0.3274039030075073
```

Output is truncated. View as a scrollable element or

Figure 9: Analysis output for the strictly confidential document in Figure 7 after masking the most influential tokens

```
"
Sample Bank AB

Ref: SB-4626-B
Document Title: Legal Assessment
Department: Compliance
Author: Anna Bergström
Date: 2025-05-07
Version: 1.4
Distribution: CFO, CTO, Relevant Department Heads

Compensation & Remuneration

The compensation committee has approved a salary review for Erik
Lund. The revised package of 1,074,789 SEK is effective from next
quarter.

Procurement Sensitive

Procurement is finalising a sole-source justification for Strata
Consulting Group AB covering services valued at 15.7 MSEK. Board
approval is required.

Commercial & Vendor Terms

Sample Bank is in exclusive negotiations with Baltic FinTech
Solutions OÜ regarding a contract extension worth up to 358.7 MSEK
over three years.
",2
```

Figure 10: Example Confidential document from synthetic dataset

Internal Memorandum: Executive Remuneration Committee Final Proposals

To: Board of Directors, Nordic Trust Bank (NTB)
From: Lars Sondergaard, Chair of the Remuneration Committee
Date: October 14, 2023
Subject: FY2023 Executive Performance Remuneration and Strategic Allocation Framework

1. Executive Summary

Following the conclusion of the third quarter and the preliminary audit of our year-to-date fiscal performance, the Remuneration Committee has finalized the proposed distribution of the Executive Bonus Pool for the 2023 fiscal year. This cycle's distribution is significantly influenced by the 18.4% increase in Net Interest Margin (NIM) achieved following the restructuring of our sovereign debt portfolio, a figure that remains restricted until the official Q4 earnings release.

The total pool has been set at €48.5 million, reflecting a 12% increase over the prior year. This reflects not only the bank's record-breaking profitability but also the successful execution of "Project North Star"—the quiet acquisition of our primary competitor's wealth management division in the Baltic region.

2. Strategic Rationale for Distribution

The Committee has moved away from traditional performance metrics to prioritize "Strategic Resilience" and "Market Expansion" milestones. The weightings for this year's payouts are structured as follows:

- **40% Operational Efficiency:** Successful integration of the AI-driven risk assessment protocols.
- **30% Regulatory Compliance:** Navigation of the revised Basel IV standards with zero identified deficiencies.
- **30% Growth & Acquisition:** Contribution toward the upcoming merger with Sjöberg & Co, which is scheduled for public announcement in January 2024.

3. Proposed Bonus Distribution Table

The following table outlines the individual allocations for the C-Suite and Senior Management. These figures include both the cash component and the deferred Restricted Stock Units (RSUs) vesting over a three-year period.

Executive Name	Designation	Performance Tier	Base Salary (EUR)	Proposed Bonus (EUR)	RSU Allocation (%)
Erik Thorsen	Chief Executive	Tier 1 -	1,200,000	4,500,000	60%

Figure 11: Example Strictly Confidential synthetic document provided by Norion Bank, not indicative of all provided documents (Page 1 of 2)

	Officer	Exceptional			
Ingrid Nilsson	Chief Financial Officer	Tier 1 - Exceptional	850,000	2,800,000	50%
Marcus Vinter	Chief Operating Officer	Tier 2 - High	720,000	1,950,000	45%
Sofia Lindholm	Chief Risk Officer	Tier 1 - Exceptional	780,000	2,400,000	50%
Anders Holm	Head of Investment Banking	Tier 2 - High	650,000	3,100,000	40%
Birgitta Svendsen	Chief Technology Officer	Tier 2 - High	680,000	1,600,000	45%
TOTAL	—	—	**4,880,000**	**16,350,000**	—

4. Special Retention Fund (Project Baltic Bridge)

In light of the upcoming Sjöberg & Co acquisition, a separate discretionary pool of €5.5 million has been established to ensure the retention of key personnel through the transition period (2024–2025). These payments are contingent upon the successful migration of €12B in assets under management without more than a 2% attrition rate.

Payments from this fund will be disbursed in two tranches:

1. **Tranche A (June 2024):** 40% upon successful regulatory approval of the merger.
2. **Tranche B (January 2025):** 60% upon the one-year anniversary of the operational integration.

5. Clawback Provisions and Risk Alignment

To maintain alignment with the European Banking Authority (EBA) guidelines, all bonuses listed above are subject to strict clawback provisions. Should any financial restatements occur within the next 24 months—specifically regarding the valuation of the non-performing loan (NPL) carve-outs in the Finnish market—the Board reserves the right to recoup up to 100% of the cash component and cancel all unvested RSUs.

6. Communication Protocol

The details contained in this document are not to be shared outside of the Board of Directors and the Group HR Lead. Any premature disclosure of these figures, particularly the NIM outperformance or the Sjöberg & Co acquisition details, would constitute a breach of fiduciary duty and could lead to market manipulation investigations by the Swedish Financial Supervisory Authority (Finansinspektionen).

Formal offer letters will be distributed via secure portal on November 15, following the internal audit sign-off.

Figure 12: Example Strictly Confidential synthetic document provided by Norion Bank, not indicative of all provided documents (Page 2 of 2)