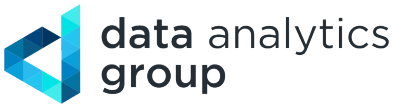

Machine Learning for Sequential Data: Unraveling the Challenges Associated with Feature Encoding, Output Decoding, and Distribution Shifts

Der Fakultät für Wirtschaftswissenschaften der

Universität Paderborn

zur Erlangung des akademischen Grades

Doktor der Wirtschaftswissenschaften

- *Doctor rerum politicarum* -

vorgelegte Dissertation

von

Matthew Caron (M.Sc.)

geboren am 22.03.1986 in St-Hyacinthe

2025

Remember that all models are wrong; the practical question is how wrong do they have to be to not be useful.

– George Box (1987)

Acknowledgements

The last few years have been an incredible journey, and I would like to take a moment to thank the many people who supported me along the way:

Oliver, for giving me this incredible opportunity, for his generous mentorship, insightful feedback, and constant support. His guidance, critical insights, and ideas have been indispensable throughout this journey.

Carina, for her constant support and exceptional organization, ensuring that everything behind the scenes ran smoothly and making our work as easy as possible.

Jochen and Michael, for the inspiring collaboration, great project ideas, and for opening my eyes to the possibilities at the intersection of data science and sports analytics – an opportunity that has shaped my career path.

Guido, for the many insightful discussions, valuable feedback, and thoughtful advice that have helped me grow academically and professionally.

David and Frederik, for always believing in me and for being incredible friends and confidants – always being there for me, day or night.

Johannes, for the great collaboration, engaging discussions, and willingness to always lend a hand. Your contributions and friendship will not be forgotten.

Last but not least, Maryna, for always being there for me, for her endless support, patience, and understanding. Knowing that you are proud of me and always by my side means everything.

Contents

List of Figures

Part A

List of Tables

Part A

Part A

Synopsis

1 Introduction

1.1 Motivation

Let's face it. Our world moves in sequences, following a natural order where one event leads to another. We wake up in the morning, follow a routine, make numerous decisions throughout the day, and react to events as they unfold. Whether commuting to work, attending a class, or simply having a conversation, our actions are shaped by what came before and influence what comes next. In short, life itself is structured around time, where past actions and choices influence current decisions and determine our future.

Like our daily lives, much of the data generated and collected today is sequential in nature. From financial transactions to online interactions or even passes in a football match, data is often recorded as a series of chronological observations. Even natural language follows a sequential order, where the meaning of a word, a sentence, or a paragraph depends on what came before. As a result, understanding and predicting sequential patterns is a key challenge in many fields, requiring models that can process data in a way that reflects its temporal structure. In fact, sequential data provides a fair representation of how things happen in our world, capturing events and processes that unfold over time rather than at a fixed moment.

However, while the field of machine learning (ML) has made significant progress in recent years, handling sequential data still presents unique challenges. In fact, unlike static data, where each observation is independent, sequential data requires models that can track patterns over time and understand how past observations influence future ones. Yet, existing methods often struggle to capture such dependencies, produce structured predictions, or maintain performance as conditions evolve.

1.2 Objectives

Building on the challenges outlined above, the primary goal of this thesis is to contribute to the body of work on sequential data modeling by addressing three key as-

pects, namely feature encoding, output decoding, and distribution shifts. By addressing these challenges, this work aims to advance how models represent sequential patterns, generate structured predictions, and adapt to changing data distributions. In doing so, it provides a basis for developing ML models that are both reliable and practical across different domains. Specifically, this thesis focuses on the following objectives:

1. *Feature encoding* – i.e., to develop strategies that allow models to capture sequential dependencies rather than relying on static representations. This includes developing encoding methods for numerical, textual, spatio-temporal, and event-based data, ensuring that models retain relationships between observations rather than treating them as isolated inputs.

2. *Output decoding* – i.e., to ensure that model predictions align with the structure and constraints of real-world problems. Instead of generating outputs as independent values, this research explores methods that maintain consistency and reliability across sequential predictions.

3. *Distribution shifts* – i.e., to develop an evaluation framework to detect, mitigate, and report issues arising from changes in data distributions over time or across different entities. This includes identifying risks such as shortcut learning and proposing solutions that help models generalize beyond their training conditions.

By tackling these challenges, this thesis contributes to both research and practice, providing methods that enhance the adaptability, robustness, and real-world applicability of ML models for sequential data.

1.3 List of Publications

To address these challenges, this thesis is composed of eight research articles, which have been published or submitted to venues in Information Systems (IS), Computer Science, ML, Operations Research, or Sports Analytics. Table 1.1 provides an overview of each publication, including type – i.e., journal or conference – ranking or score – i.e., VHB-Rating, CORE, and h-5 index – and citation details. Together, these publications highlight the scope and impact of the research contributions, demonstrating their alignment with the thesis objectives.

Table 1.1: List of Publications

	Publication	Type	VHB	CORE	h-5 index
Paper 1	Caron, M. and Müller, O. (2020). Hardening Soft Information: A Transformer-Based Approach to Forecasting Stock Return Volatility. In *Proceedings of the IEEE International Conference on Big Data*, pages 4383–4391.	Conference	–	B	54
Paper 2	Müller, O., Caron, M., Döring, M., Heuwinkel, T., and Baumeister, J. (2021). PIVOT: A Parsimonious End-to-End Learning Framework for Valuing Player Actions in Handball using Tracking Data. In *Proceedings of Workshop on Machine Learning and Data Mining for Sports Analytics (ECML PKDD)*, pages 116–128.	Conference	–	–	–
Paper 3	Caron, M., Gulenko, M., and Müller, O. (2021). To the Moon! Analyzing the Community of "Degenerates" Engaged in the Surge of the GME Stock. In *Proceedings of the International Conference on Information Systems*, pages 2432–2448.*	Conference	A	–	–
Paper 4	Caron, M., Bartelheimer, C., and Müller, O. (2022). Towards a Reliable & Transparent Approach to Data-Driven Brand Valuation. In *Proceedings of the Americas Conference on Information Systems*, pages 1353–1363.	Conference	C	–	–
Paper 5	Caron, M. (2022). Shortcut Learning in Financial Text Mining: Exposing the Overly Optimistic Performance Estimates of Text Classification Models under Distribution Shift. In *Proceedings of the IEEE International Conference on Big Data*, pages 3486–3495.	Conference	–	B	54
Paper 6	Dieter, P., Caron, M., and Schryen, G. (2023). Integrating Driver Behavior into Last-Mile Delivery Routing: Combining Machine Learning and Optimization in a Hybrid Decision Support Framework. *European Journal of Operational Research*, 311(1).	Journal	A	–	117
Paper 7	Caron, M. and Müller, O. (2023). TacticalGPT: Uncovering the Potential of LLMs for Predicting Tactical Decisions in Professional Football. In *Proceedings of the StatsBomb Conference*.	Conference	–	–	–
Paper 8	Caron, M., Müller, O., and Kriebel, J. (2025). Detecting and Mitigating Shortcut Learning Bias in Machine Learning: A Pathway to More Generalizable ML-based (IS) Research. *Working Paper Series, Paderborn University, Faculty of Business Administration and Economics*, (129).**	Journal	(A+)	(–)	(60)

* *The paper was nominated for Best Paper at the International Conference on Information Systems (ICIS) 2021.*
** *The manuscript was submitted for publication at Information Systems Research in February 2025. An earlier version of this manuscript was also presented at the Workshop on Information Technologies and Systems (WITS) 2022.*

1.4 Thesis Structure

This thesis is organized into two main parts – i.e., **Part A – Synopsis** and **Part B – Research Papers** – which provide a structured overview of the research, covering the methodological contributions and their practical applications.

Part A – Synopsis outlines the research background, key challenges, and the contributions made in this thesis and consists of the following chapters:

- *Chapter 1 – Introduction:* Introduces the motivation, research objectives, and structure of the thesis.
- *Chapter 2 – Research Background:* Reviews relevant literature on sequential data modeling and the key challenges in feature encoding, output decoding, and distribution shifts.
- *Chapter 3 – Research Contributions:* Summarizes the main findings from the research papers and how they address the identified challenges.
- *Chapter 4 – Discussion & Conclusion:* Discusses the broader implications of the research, its limitations, and future directions.

Part B – Research Papers contains the eight research articles that form this thesis. These papers collectively demonstrate the application of the methodologies discussed in **Part A** to real-world problems in domains such as finance, branding, logistics, and sports analytics. Each paper addresses specific methodological challenges while contributing to the goal of improving research in the field of sequential data modeling.

2 Research Background

2.1 Overview

Understanding sequential data is critical in many ML problems where observations are not independent but evolve over time. Unlike static datasets, where standard approaches can be applied without considering order, sequential data requires models that capture dependencies between past and future observations. As Hausman explains, "the key difference between sequential and non-sequential decision problems is that future decisions in sequential problems may be based partially on information known in the future but unknown at present" (1969, p. B-93), meaning that later decisions may not only depend on earlier ones but also on information that becomes available as the sequence unfolds (Hausman, 1969). For example, in weather forecasting, predictions are continuously updated as new temperature and pressure readings are recorded over time. Similarly, Wittenbach et al. (2020) argue that temporal data, such as financial sequences or transactional records, presents unique patterns and dependencies that cannot be effectively captured using traditional modeling approaches. They also emphasize the need for specialized methods to handle these time-dependent structures, as static models often fail to fully leverage the complex relationships within sequential datasets.

Hence, this thesis focuses on sequential data because learning from these structures introduces fundamental challenges that impact how data is encoded, dependencies are captured, and outputs are generated reliably. Since these challenges vary depending on the data type, we first define the most common forms before addressing the key difficulties in modeling them.

2.2 Sequential Data

2.2.1 Definition

As briefly exposed, sequential data consists of ordered observations where past values influence future ones. In traditional supervised learning, where the goal is to train a

model that maps inputs to outputs based on labeled examples, the standard assumption is that the training data is independent and identically distributed (i.i.d.) – i.e., each input-output pair is drawn randomly from a joint distribution (Dietterich, 2002; Raschka and Mirjalili, 2017). This assumption simplifies modeling by ensuring that each observation is treated independently, allowing standard ML methods to generalize well across different scenarios (Dietterich, 2002).

However, as Dietterich (2002) explains, sequential data does not fit this assumption, as observations exhibit strong dependencies across time. Instead of being randomly sampled, data points in sequences are linked by patterns that evolve over time, making traditional learning techniques inadequate. This is especially relevant in applications where the temporal order is crucial (Dietterich, 2002; Raschka and Mirjalili, 2017). For example, in stock price forecasting, the value of a share at a given moment is usually closely related to previous prices. Similarly, in text classification, the meaning of a word often depends on the words that came before it. In such cases, predictions must consider how patterns develop over time rather than treating each observation as separate.

As a result, sequential data cannot be modeled in the same way as static data. Instead, predictions in sequential modeling must be made by considering the dependencies between observations rather than treating them in isolation (Dietterich, 2002). Such dependencies can be expressed as a sequence $(x_1, x_2, \ldots, x_T)$, where x_t represents an observation at step t, and T is the total sequence length (Dietterich, 2002). These dependencies can be short-term – i.e., where only recent observations matter – or long-term – i.e., where relationships span across distant elements in the sequence.

To better understand these dependencies, the following subsections introduce the most common types of sequential data and their defining characteristics.

2.2.2 Time Series Data

A time series is a collection of observations recorded in temporal order, where each data point is indexed according to the sequence in which it occurs (Box et al., 2008; Shumway and Stoffer, 2025). Therefore, it can be defined as a stochastic process, represented as a collection of random variables x_t indexed by time t (Shumway and Stoffer, 2025, p. 10). Unlike datasets with independent observations, time series data exhibits serial dependence, meaning that past values influence future ones (Box et al., 2008; Shumway and Stoffer, 2025). Technically, "[a]nything that is observed sequentially over time is a time series" (Hyndman and Athanasopoulos, 2018, p. 17), meaning that such data may

be recorded at fixed (e.g., hourly, daily, or annually) or irregular intervals, depending on the application (Wittenbach et al., 2020). Since time series analysis typically focuses on tracking the evolution of a single entity over time, it requires specialized techniques that account for sequential dependencies. Hence, in time series forecasting, the goal is to estimate how a sequence will evolve based on past observations, often relying on patterns in the data rather than external factors (Hyndman and Athanasopoulos, 2018).

The structure of a time series depends on how observations are recorded over time. If data points are collected at fixed, evenly spaced intervals, the series is referred to as regular or uniform (Wittenbach et al., 2020). This is the case for hourly stock prices or daily energy consumption measurements. In contrast, when the timing of observations varies, the series is considered irregular or non-uniform, where observations are recorded dynamically based on external events rather than a predefined schedule (Wittenbach et al., 2020). For example, financial transactions such as credit card payments or player actions during a football match do not follow a fixed schedule but depend on individual decisions or game dynamics. While many ML approaches attempt to transform such data into a uniform format, doing so can result in information loss, as the timing of events may hold predictive value (Wittenbach et al., 2020). This category of non-uniform time series, often referred to as event data or non-uniform event streams, has distinct temporal and structural properties, which will be discussed in a later section.

2.2.3 Panel Data

In contrast to time series data, which focuses on a single entity over time, panel data consists of repeated observations collected for multiple entities over a given period (Verbeek, 2004; Wooldridge, 2013). These entities can be, for instance, individuals, firms, or even geographic regions, each observed across time (Verbeek, 2004). Structurally, panel data extends the principles of time series by capturing multiple time series simultaneously – i.e., one for each entity in the dataset (Wooldridge, 2013). Unlike a standard cross-sectional dataset, which captures observations at a single point in time, panel data incorporates a temporal dimension by tracking multiple entities over time, combining cross-sectional and time-series properties (Verbeek, 2004). This structure enables the capture of differences between entities and variations within the same entity over time, allowing for more detailed analyses (Verbeek, 2004). However, panel data introduces complexities because, unlike standard datasets, we cannot assume that observations are independent across time or entities, requiring specialized statistical methods to account for temporal and entity-level dependencies (Verbeek, 2004; Wooldridge, 2013).

Panel data is widely used in empirical research, particularly when analyzing trends at an individual or organizational level. Since it consists of multiple time series, panel data is especially useful when comparing patterns across different entities. Typical applications include financial analyses, where investments and performance indicators are monitored over time. Similarly, it is applied in sustainability and corporate social responsibility research, where firms' environmental and social efforts are systematically recorded to assess their long-term impact on business and the environment.

2.2.4 Event Data

As explained above, non-uniform time series, also known as event data or event streams, differ from traditional time series in that observations occur at irregular intervals without a predefined sampling frequency (Wittenbach et al., 2020). While uniform time series record observations at fixed time steps, event data consists of discrete occurrences associated with a timestamp rather than continuous measurements (Wittenbach et al., 2020). Unlike panel data, where repeated observations are recorded uniformly for multiple entities over time, event data focuses on the order and timing of specific events rather than maintaining regular intervals. Still, even though irregular intervals between events can carry meaningful information, many ML methods attempt to convert event data into uniform time steps, often losing important details (Wittenbach et al., 2020).

Contrary to time series, where observations follow a fixed schedule, one key characteristic of event data is that the exact timing of events carries important information (Wittenbach et al., 2020). For example, online purchases, social media posts, or business processes, such as order placements or product shipments, are all examples of event data since these events occur in response to specific situations rather than at regular intervals. In fact, the field of Business Process Management (BPM) – i.e., a key area of IS research – relies heavily on event-driven data, where business activities are recorded as irregular occurrences. Similarly, as exposed earlier, player actions in football, such as shots, passes, or tackles, also qualify as event data since they happen in response to game situations rather than a set schedule. As a result, event data is widely used in football analytics to assess player performance and tactical strategies (Decroos et al., 2019). For example, Anzer and Bauer (2021) highlight that shot events contain detailed attributes describing various aspects of the action, such as shot type or defensive pressure, which are crucial for evaluating player actions. Generally speaking, such contextual factors are essential for analyzing event patterns, identifying trends, and estimating probabilities, allowing models to extract deeper insights from event data.

2.2.5 Spatio-Temporal Data

As introduced in the previous sections, datasets with a temporal component can take different forms depending on how observations are organized across entities and time. While time series focus on a single entity over time, and panel data track multiple entities across time, spatio-temporal data, sometimes referred to as tracking data, incorporates an additional spatial dimension. In other words, spatio-temporal data focuses on observations "related to each other in the context of space and time" (Atluri et al., 2018, p. 83:2). This structure enables the study of patterns that evolve across both spatial and temporal dimensions (Moraga, 2024), such as player movements on a football pitch, the flow of traffic in a busy city, or the spread of an infectious disease across geographic regions. Given the increasing availability of large-scale data, spatio-temporal analysis has become essential in various fields, including climate science, social sciences, neuroscience, epidemiology, and transportation, where understanding interactions across space and time is crucial (Atluri et al., 2018).

Similar to time series data, spatio-temporal data can be collected in different ways, leading to variations in how the observations are structured. Some datasets are uniform – i.e., recorded at fixed spatial locations and time intervals – while others are non-uniform – i.e., recorded at varying locations and time intervals (Atluri et al., 2018; Moraga, 2024). For instance, weather stations measuring temperature at exact locations every hour form a uniform spatio-temporal dataset. In contrast, event data with spatial references, such as shots or passes in football, result in non-uniform spatio-temporal data, as these actions occur irregularly throughout a match. Hence, tracking both the timing and location of such events enables analysis of player movement and tactical positioning (Atluri et al., 2018; Anzer and Bauer, 2021). These differences affect how spatio-temporal data is analyzed, as uniform datasets support traditional statistical modeling, whereas non-uniform data often requires interpolation or specialized ML techniques, highlighting the limitations of i.i.d. assumptions (Atluri et al., 2018).

2.2.6 Textual Data

Lastly, one of the most widely used but perhaps less obvious forms of sequential data is textual data, also known as natural language. While language is often seen as nothing more than discrete words or characters, its structure is inherently sequential, where the meaning of each word, sentence, or paragraph depends on its position within a broader context. In fact, "[t]he ordering of words becomes important as they convey logical

relationships and dependencies between the words" (Patil et al., 2023, p. 36124), giving a document its meaning, as rearranging them would completely change interpretation.

That being said, language is complex to understand and analyze – i.e., even for humans – as it involves intricate dependencies and ambiguities that shape meaning and context. From a technical perspective, natural language processing – i.e., "[...] the subfield of computer science concerned with using computational techniques to learn, understand, and produce human language content" (Hirschberg and Manning, 2015, p. 261) – provides the foundation for modeling and analyzing these complexities.

Unlike time series, panel, event, or spatio-temporal data, natural language sequences are ordered but do not possess a conventional temporal structure. Instead, language follows a strict sequence where order determines meaning, even though there is no inherent concept of time between tokens. In NLP, each time step corresponds to a token (e.g., a character or a word) rather than a measurement at a specific moment.

To process these sequences computationally, tokens are transformed into numerical representations. A basic approach is one-hot encoding – i.e., assigning each token a unique binary vector where only one position is active. However, this representation does not capture relationships between words. In contrast, more recent methods, such as RNNs or Transformers, process sequences by representing each token as an embedding at each time step. These embeddings are dense vector representations that capture semantic relationships between words, allowing models to learn contextual dependencies and distinguish meanings (Patil et al., 2023). Such techniques have become standard in modern NLP, driving applications from machine translation to AI assistants.

2.3 Challenges of Modeling Sequential Data

2.3.1 Overview

As established in the previous section, sequential data differs fundamentally from static datasets, as observations are connected across time and, in some cases, space. Unlike independently sampled data points, sequential data requires ML models that can recognize and learn from these dependencies, whether they emerge from temporal patterns, spatial-temporal relationships, or text documents. Additionally, sequential data exhibits varying levels of dependency, ranging from short-term relationships – i.e., where only recent observations matter – to long-term structures – i.e., where patterns evolve over extended sequences. These difficulties not only make many traditional modeling

techniques unsuitable but also introduce challenges in how the data needs to be encoded – i.e., represented – and decoded. Moreover, sequential data is prone to distribution shifts over time, as external factors, changing conditions, or evolving situations modify its statistical properties, making it difficult for models to generalize effectively.

Hence, in this section, we explore three key challenges in modeling sequential data, namely feature encoding, output decoding, and distribution shifts. Each of these challenges plays a critical role in determining how models learn from sequences, influencing both their predictive performance and generalization ability. These challenges are central to the research contributions presented in this thesis and directly shape the methods proposed in the next chapters.

2.3.2 Feature Encoding & Output Decoding

At this point, it should be clear that transforming sequential data into structured representations is, even though implicit, a fundamental challenge in machine learning that should not be overlooked. As a matter of fact, several characteristics of sequential data make feature encoding and output decoding particularly complex, requiring methods that can preserve structure, capture dependencies, and ensure reliable model outputs. These challenges arise from the nature of sequential data itself and can be broken down into the following key aspects:

- *Temporal Dependencies:* Unlike static datasets, where each observation is independent, sequential data requires encoding methods that preserve relationships between consecutive observations, such as Recurrent Neural Networks (RNN) (Sherstinsky, 2020). Yet, the challenge lies in ensuring that short-term fluctuations and long-term trends are accurately captured, which is particularly important in panel data, time series forecasting, and NLP. Long Short-Term Memory (LSTM) networks (Sherstinsky, 2020) and, more recently, Transformer-based models (Vaswani et al., 2017), such as BERT (Devlin et al., 2019), are commonly used to address this challenge, as they are designed to retain both short-term dependencies and long-range patterns.

- *Structural Complexity and Entity Interactions:* In panel data and event data, relationships exist not just across time but also between different entities (e.g., athletes, companies, or geographical regions). Therefore, encoding strategies must account for hierarchical dependencies, entity-specific variations, and interactions, which complicate how models learn from such data.

- *Irregular Sampling and Non-Uniformity:* As exposed, sequential data can be recorded at irregular intervals, as seen with event or spatio-temporal data, such as football match events where player actions occur unpredictably throughout the game (Decroos et al., 2019). Hence, encoding methods must be able to handle missing or unevenly spaced observations without distorting the underlying patterns. Standard techniques assuming uniform time steps may fail in these cases, leading to information loss.

- *High Dimensionality and Multi-Modality:* Sequential datasets often combine textual, numerical, and categorical information, creating what is known as multimodal data. Encoding such data requires methods that can not only represent each modality but also capture their dependencies effectively rather than treating them as independent inputs (Zhao et al., 2017).

- *Variable-Length Outputs and Structured Predictions:* Many sequential models generate complex outputs, such as spatio-temporal forecasts, event sequences, or textual data. A key challenge arises when these outputs cannot be represented with fixed-dimensional vectors, since "many important problems are best expressed with sequences whose lengths are not known a-priori " (Sutskever, 2014, p. 1). Decoding such outputs requires methods that ensure consistency, preserve structural dependencies, and adapt to variations in sequential patterns so that generated sequences remain reliable and aligned with the learning task.

The challenges outlined in this section highlight the difficulties of modeling sequential data, particularly in how features are encoded and outputs are structured. Effective solutions must be able to capture evolving dependencies, preserve structure, and handle variability in sequential patterns. In the next chapters, we introduce the research contributions of this thesis, each addressing one or more of these challenges with methods designed for the specific properties of the sequential data examined in each paper.

2.3.3 Distribution Shifts in Machine Learning

As discussed earlier, ML models are typically trained under the assumption that training and test data share the same distribution. However, this assumption of independent and identically distributed (i.i.d.) data rarely holds in practice and has been referred to as the "big lie of machine learning" (Varshney, 2022, p. 114). In practice, the distribution of data a model encounters during deployment often differs from the training data, a phenomenon known as distribution shift (Varshney, 2022). When this oc-

curs, models that perform well in controlled environments may struggle in real-world or post-deployment scenarios, leading to unreliable predictions and decreased performance (Varshney, 2022; Kulinski and Inouye, 2023).

Technically speaking, distribution shifts can be represented as

$$p_{X,Y}^{(\text{train})}(x, y) \neq p_{X,Y}^{(\text{deploy})}(x, y)$$

where the joint distribution of features and labels differs between training and deployment (Varshney, 2022). This mismatch depends on which part of the distribution changes and is commonly classified into three main types (Varshney, 2022):

- *Label shift*, also known as *prior probability shift*, occurs when the distribution of labels changes between training and deployment, but the relationship between features and labels remains the same. In other words, while the probability of different labels varies, the features associated with each label do not change:

$$p_{Y}^{(\text{train})}(y) \neq p_{Y}^{(\text{deploy})}(y), \quad p_{X|Y}^{(\text{train})}(x|y) = p_{X|Y}^{(\text{deploy})}(x|y)$$

- *Covariate shift* happens when the distribution of features differs between training and deployment, even though the relationship between inputs and outputs remains stable. This typically occurs when data is collected under different conditions, affecting input characteristics while preserving labels:

$$p_{X}^{(\text{train})}(x) \neq p_{X}^{(\text{deploy})}(x), \quad p_{Y|X}^{(\text{train})}(y|x) = p_{Y|X}^{(\text{deploy})}(y|x)$$

- *Concept drift* refers to cases where the relationship between inputs and outputs itself changes over time, meaning that the same features may correspond to different labels in training and deployment. It can also occur when the way features relate to labels evolves, even if the label distribution remains unchanged:

$$p_{Y|X}^{(\text{train})}(y|x) \neq p_{Y|X}^{(\text{deploy})}(y|x)$$

  In some cases, concept drift manifests through shifts in the conditional distribution of features given labels, while the overall label distribution stays the same:

$$p_{X|Y}^{(\text{train})}(x|y) \neq p_{X|Y}^{(\text{deploy})}(x|y), \quad p_{Y}^{(\text{train})}(y) = p_{Y}^{(\text{deploy})}(y)$$

To put this into perspective, consider the following examples. *Label shift* occurs in loan approvals if the proportion of borrowers who default and those who do not changes over time, even though the criteria used to assess risk remain the same. *Covariate shift* happens in a housing price prediction model trained on urban homes but later used in rural areas – i.e., where houses differ in sizes and locations – but the relationship between features and prices remains unchanged. *Concept drift* occurs in email spam detection when the way emails are composed and formatted changes, making the language that once indicated spam no longer a reliable signal.

In conclusion, distribution shifts can severely impact model performance, causing predictions to become unreliable and leading to significant degradation in real-world applications. Therefore, identifying and addressing these shifts is essential for building reliable ML systems. While detection methods compare distributions between training and deployment data to identify shifts, mitigation strategies focus on adapting models to maintain performance under changing conditions (Varshney, 2022). The impact of these shifts also depends on the type of sequential data. For instance, concept drift is particularly problematic in financial forecasting, where market dynamics change over time, while covariate shift is common in spatio-temporal data, such as traffic prediction, where new urban developments alter input distributions. Hence, ensuring model performance and reliability requires effective identification and mitigation strategies, which we also address in the next chapters.

3 Research Contributions

3.1 Overview

As outlined in Table 3.1, the eight research contributions presented in this thesis address, as a whole, the main challenges associated with sequential data modeling. They cover a diverse range of data types – i.e., multimodal, spatio-temporal, textual, event, and numerical – each posing distinct methodological challenges related to feature encoding, output decoding, and distribution shifts. This table provides an overview of how these contributions relate to key challenges in ML.

A recurring aspect across these studies is the need to capture sequential dependencies at both the feature level, where relationships exist within the input variables, and the observation level, where past observations influence future predictions. In contrast, textual and event data present different challenges, requiring models to account for sequential order and entity interactions. Distribution shifts further complicate these tasks, as models often struggle when data distributions change over time or across entities. Throughout this thesis, we examine these difficulties, introduce structured encoding strategies, and develop modeling approaches tailored to the complexities of sequential data. As outlined below, these contributions present frameworks and methods that improve how models encode, learn from, and generate structured, sequential outputs while addressing distribution shifts.

Focussing on these challenges, the contributions are structured as follows:

1. **Feature Encoding**

    - **Paper 1**, **Paper 3**, and **Paper 4** explore how multimodal feature encoding integrates textual, numerical, and categorical inputs while preserving key relationships.

    - **Paper 1**, **Paper 3**, **Paper 4**, **Paper 5**, and **Paper 8** examine panel data structures, focusing on how entity-level dependencies change over time and influence model performance.

- **Paper 2** and **Paper 6** focus on spatio-temporal encoding, where movement patterns and feature relationships change over time. These studies examine how models can preserve these evolving structures, ensuring that spatial and temporal dependencies are accurately captured during learning.
- **Paper 5** and **Paper 7** examine encoding strategies for textual and event-based data, focusing on preserving sequence structures and capturing interactions between entities. These studies explore methods to ensure that models retain the order of events and relationships within the data, improving how sequential patterns are processed and understood.

2. **Output Decoding**

   - **Paper 2** and **Paper 6** tackle challenges in spatio-temporal predictions, ensuring that model outputs remain realistic, respect domain constraints, and accurately reflect movement patterns.
   - **Paper 7** explores how generative models decode event-based sequences into structured outputs while preserving consistency and the order of events.

3. **Distribution Shifts**

   - **Paper 5** and **Paper 8** examine distribution shifts, showing how models struggle when data distributions change over time or across entities, leading to overoptimistic predictive performance.
   - **Paper 8** introduces a framework to detect, mitigate, and report shortcut learning – i.e., when models rely on spurious associations instead of meaningful patterns – leading to poor generalization under distribution shifts. This framework helps improve model evaluation by identifying weaknesses that standard assessments might overlook.

Together, these contributions provide a structured view of sequential data modeling, addressing challenges associated with feature encoding, output decoding, and distribution shifts across different learning tasks. Each paper focuses on specific methodological gaps, demonstrating how various modeling approaches can be applied to structured, unstructured, spatio-temporal, and event-based data. The following sections summarize each contribution, highlighting how they tackle key issues such as multimodal integration, temporal dependencies, model robustness, and structured output generation.

Table 3.1: Detailed Summary of the Research Contributions

	Data		**Sequential Dependencies**		**Challenges**		
	Type	*Panel Structure*	*Features*	*Observations*	*Feature Encoding*	*Output Decoding*	*Distribution Shifts*
Paper 1	Numerical, Textual	✓	✓	✓	✓	✗	(✗)**
Paper 2	Spatio-Temporal	✗	✗	✓	✓	✗	(✗)**
Paper 3	Categorical, Numerical, Textual	✓	✗	✓	✓	✗	(✗)**
Paper 4	Categorical, Numerical, Textual	✓	✗	✓	✓	✗	(✗)**
Paper 5	Textual	✓	✓	✓	✗	✗	✓
Paper 6	Spatio-Temporal	✗	✓	✗	✓	✓	✗
Paper 7	Event	✗	(✗)*	✓	✓	✓	(✗)**
Paper 8	Numerical	✓	✗	✓	✗	✗	✓

* *The features only exhibit sequential dependencies after being transformed into textual format.*

** *Instances where distribution shifts are likely to affect the learning problem despite not being addressed by the paper explicitly.*

3.2 Paper 1 – Hardening Soft Information: A Transformer-Based Approach to Forecasting Stock Return Volatility

Paper 1 investigates the challenges associated with feature encoding in sequential data, focusing on effectively processing and representing long textual documents for predictive modeling tasks. Unlike structured numerical data, textual information exhibits complex dependencies across varying lengths, making encoding difficult without losing essential context. As with most problems involving sequential data, long-range dependencies are essential for understanding and processing the connection between events over time, or in this case, concepts or phrases throughout a document. However, state-of-the-art approaches, such as Transformer architectures, struggle with feature-level dependencies since self-attention – i.e., the mechanism these architectures rely on to capture relationships between input features – scales quadratically with sequence length (Devlin et al., 2019; Beltagy et al., 2020). This computational constraint limits the maximum feasible document length these models can realistically process, posing a fundamental challenge for applying them to real-world scenarios where textual documents tend to be lengthy.

Given this background, this study applies modern attention-based sequence-to-sequence models to the regression learning task of stock return volatility prediction – i.e., a particularly realistic scenario for assessing the adaptability of Transformer models to the aforementioned challenge. More specifically, we investigate whether leveraging textual features extracted from corporate financial reports can improve predictions of stock return volatility – i.e., a standard measure of risk in finance. Historically, financial forecasting has predominantly relied on numerical, hard information, such as balance sheets or stock prices, to support decision-making. However, following advancements in natural language understanding, integrating soft information, such as textual data from corporate annual reports, into predictive models has become increasingly common in finance and IS. Hence, more than a decade after the seminal work by Kogan et al. (2009), this study revisits this task using a state-of-the-art Transformer-based approach.

Following the approach illustrated in Figure 3.1, this study introduces a feature-based method for encoding long documents while preserving contextual relationships. The framework utilizes pre-trained Transformer models, such as BERT (Devlin et al., 2019), to extract contextual embeddings from corporate annual reports, specifically, the Management's Discussion and Analysis (MD&A) section of Form 10-K filings. The process begins by tokenizing the MD&A text and segmenting it into n overlapping chunks of a predefined length m. Each chunk is then passed through the Transformer model, and

the activations from the last four hidden layers are extracted, pooled, and concatenated to form a single, feature vector representing the entire document. By leveraging mean pooling across these representations, this approach allows for efficient encoding of arbitrarily long sequences while maintaining key contextual dependencies, ensuring that essential information is not lost due to sequence length constraints.

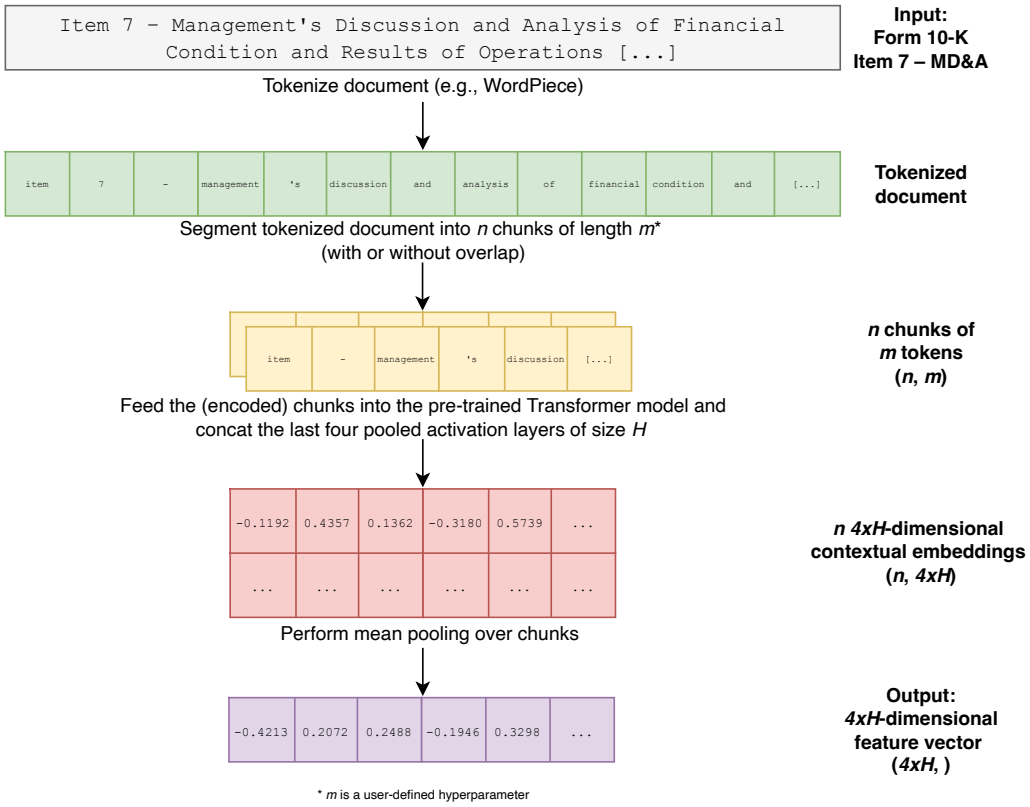

Figure 3.1: Feature-Based Approach (Caron and Müller, 2020, p.4385)

This feature-based encoding method offers a scalable solution to handling long financial texts, overcoming the standard limitations of Transformer architectures with fixed-length inputs. By preserving essential textual patterns and contextual information, it enables forecasting models to incorporate soft information alongside traditional numerical indicators. This capability is particularly valuable in real-world financial applications, where corporate disclosures often exceed standard sequence lengths and contain critical insights that affect market expectations and risk assessments.

For our empirical evaluation, we used the publicly available FIN10K dataset, which includes corporate annual reports from 1996 to 2013 (Tsai et al., 2016). Our experiments tested text-only and multimodal configurations, where textual embeddings were used alone or combined with structured numerical features. As shown in Table 3.2,

our models outperformed prior benchmarks, demonstrating the effectiveness of incorporating textual data into financial forecasting. The text-only model achieved lower prediction errors than traditional baselines relying solely on numerical information, confirming the predictive value of soft financial information. However, the best results were obtained using a combined approach, which integrated both textual and numerical features, achieving the lowest mean squared error across multiple years. These findings reinforce the importance of leveraging diverse data sources in sequential financial modeling, showing that financial forecasting models benefit from a richer representation of firm-specific risk factors when incorporating textual insights alongside structured data.

Table 3.2: Best Models vs. Published Results (Caron and Müller, 2020, p.4389)

Model	2008	2009	2010	2011	2012	2013	AVG
Hist. vol. (baseline)	0.4872	**0.2065**	0.1858	0.0802	0.1508	0.0796	0.1984
Tsai et al. 2016 (EXP-SYN)	0.6537	0.2387	0.1514	0.1217	0.2290	0.1861	0.2634
Dereli and Saraçlar 2019 (CNN-NTC)	0.4672	0.3169	0.2156	0.1154	0.1944	0.1238	0.2389
Ours (text-only)	**0.3241**	0.2672	0.1383	0.0964	0.1423	0.1007	0.1782
Ours (combined)	0.3801	0.2170	**0.1366**	**0.0733**	**0.1302**	**0.0720**	**0.1682**

** The above performance results are displayed in terms of mean squared error*

3.3 Paper 2 – PIVOT: A Parsimonious End-to-End Learning Framework for Valuing Player Actions in Handball using Tracking Data

Modeling high-frequency spatio-temporal data presents fundamental challenges when it comes to data representation and feature encoding, particularly when statistical properties shift over time and space. Unlike structured datasets with well-defined attributes, spatio-temporal sequences exhibit heterogeneity and non-stationarity, where the relationships between observations evolve dynamically. This variability complicates predictive modeling, as traditional approaches often rely on fixed representations that fail to generalize across changing contexts.

To explore these challenges, **Paper 2** proposes an end-to-end learning framework designed for estimating expected possession value (EPV) in professional handball – i.e., a predictive modeling task that assesses the impact of in-game actions on scoring likelihood. While EPV estimation has been widely studied in sports such as basketball and football, prior methods often rely on predefined event annotations, making them

less adaptable to sports or scenarios with continuous, fast-paced play. Instead, this framework encodes continuous spatio-temporal data without requiring predefined event labels, allowing for a more flexible and scalable approach that generalizes beyond handball to other domains characterized by continuous spatio-temporal dynamics.

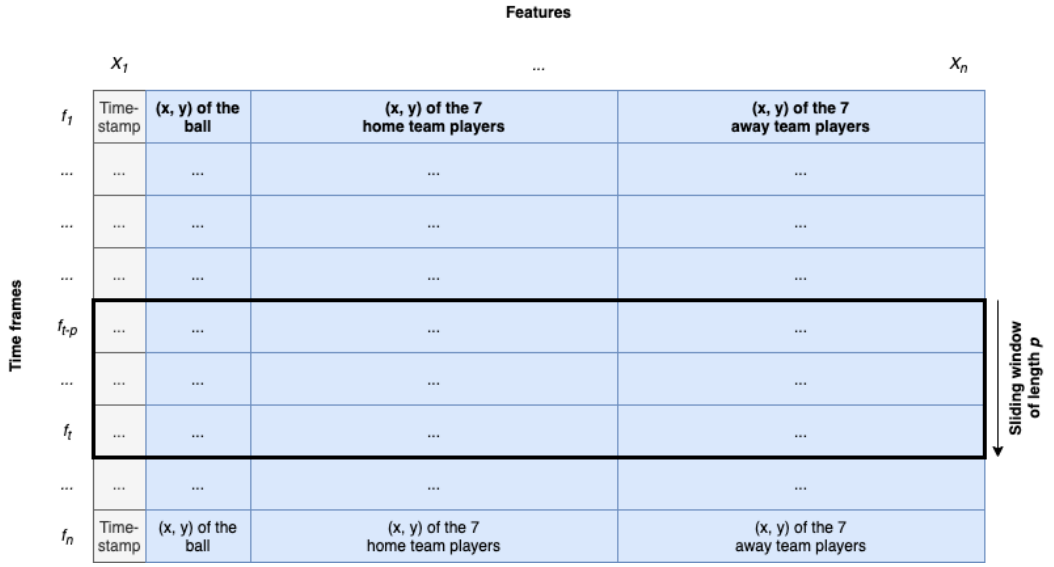

Figure 3.2: Spatio-Temporal Representation of Game Dynamics (Müller et al., 2021, p.119)

Following this framework, the study represents spatio-temporal game dynamics as structured input sequences, enabling learning algorithms to extract meaningful patterns directly from raw tracking data. As shown in Figure 3.2, player and ball positions are encoded as a sequence of two-dimensional spatial arrays, capturing movement over a fixed time window. This approach ensures that models process continuous player trajectories and positional interactions without relying on labeled events, maintaining adaptability to different styles of play and game contexts. The learning task is formulated as a binary classification problem, where the model predicts the probability of the attacking team scoring within a predefined number of frames, providing an interpretable measure of how in-game actions influence possession value over time.

To evaluate the framework, we used tracking data from the 2019/20 season of the Liqui Moly Hanbdall-Bundesliga, ensuring that the dataset was structured into fixed-length input sequences using a sliding window technique. The evaluation compared multiple deep learning architectures, including Fully Convolutional Networks (FCN), Long Short-Term Memory (LSTM) networks, and Time Series Transformers (TST). Each model was trained to predict the scoring likelihood within the next 3 seconds – i.e., 60 frames at 20 frames per second – based on spatial game dynamics alone, allowing for

a direct assessment of how well different architectures capture spatio-temporal dependencies in high-frequency data. Our results show that the TST model outperformed all other approaches across all tested window lengths, demonstrating its superior ability to model complex movement patterns. Notably, the TST model achieved an AUC of 0.909 and a Brier Skill Score of 0.318, highlighting its effectiveness in forecasting possession outcomes based purely on tracking data.

Beyond predictive accuracy, the study also explored practical applications of the framework to real-time tactical analysis. As demonstrated in Figure 3.3, the model enables the development of an *Augmented Instant Replay* system, which continuously updates EPV estimates throughout a game, providing a visual representation of how team actions influence scoring probabilities. This system offers potential applications for the coaching staff, analysts, and broadcasters, allowing for a real-time data-driven assessment of tactical decisions. For instance, during an attack by SG Flensburg-Handewitt in a league match, the model captured a steep rise in EPV following a fast break initiated by the goalkeeper, illustrating how possession value fluctuates dynamically based on game events.

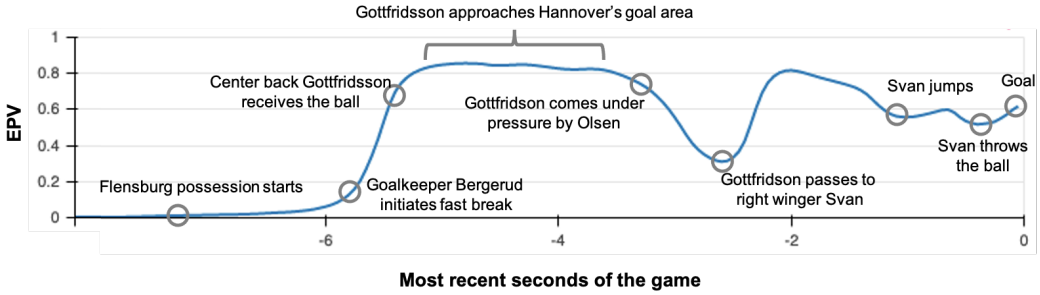

Figure 3.3: EPV Development of an Attack by SGFH (Müller et al., 2021, p.125)

3.4 Paper 3 – To the Moon! Analyzing the Community of "Degenerates" Engaged in the Surge of the GME Stock

Addressing the question of how multimodal data can improve modeling outcomes, **Paper 3** examines the challenges associated with representing complex multimodal panel feature sets – i.e., feature sets comprised of textual, numerical, and time-series data – within a modeling approach suited to address common challenges encountered in IS and finance research. Unlike structured datasets with well-defined attributes, multimodal data must be carefully encoded to ensure that each data type is effectively integrated, allowing diverse inputs to contribute meaningfully to predictive performance.

To give our study a current real-world context, we focused, in **Paper 3**, on the events surrounding the dramatic price surge of the GameStop Corp. (GME) stock of 2021, analyzing the relationship between user activity on the *r/wallstreetbets* (WSB) subreddit and the trading volume of the GME stock. This case presents a natural setting for investigating the challenges of modeling multimodal data, as irregular user activity patterns and the mix of structured and unstructured inputs make it more difficult to capture meaningful relationships across different data types.

To recall the events of early 2021, the financial world was taken by storm when, over the course of just four weeks, the price of the GME stock skyrocketed by an extraordinary 2,442% (see Figure 3.4) – i.e., a rise attributed, in part, to the actions of the now-famous WSB subreddit group. During this time, retail investors on WSB organized what became one of the most widely publicized short squeezes in history, drawing attention not only to GameStop but also to other stocks like AMC Entertainment Holdings, Inc. (AMC) and BlackBerry Limited (BB), which experienced similar surges. These events were described as unprecedented cases of predatory trading by retail investors, exposing the reach and impact of so-called "kitchen table trading" (Hasso et al., 2022).

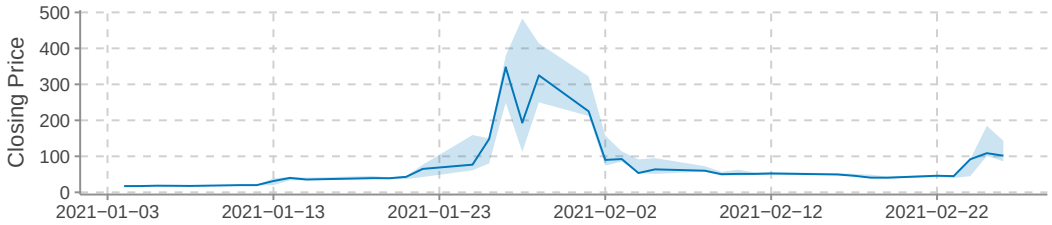

Figure 3.4: Daily Closing Price of GameStop Corp. in USD (Caron et al., 2021, p.2)

For the analysis, we constructed a dataset spanning 105 trading days, covering the period from August 1, 2020, to March 15, 2021. The dataset included over 169,000 posts, 4.78 million comments, and detailed financial market data, from which we derived more than 300 predictors capturing various aspects of user activity. To structure this multimodal data for predictive modeling, we applied a feature encoding process that transformed textual discussions into numerical representations while ensuring alignment with time-series financial indicators. To proceed, we first extracted every document's sentiment, or polarity, using a pre-trained text classification model based on the architecture by Liu et al. (2019). Then, we applied Latent Dirichlet Allocation (LDA) to extract discussion themes, identifying recurring topics in the documents. These extracted features were then synchronized with daily trading data, ensuring that the textual and financial components remained temporally aligned for effective modeling.

With the feature set constructed, we employed four Bayesian Structural Time Series (BSTS) models – i.e., models that integrate temporal trends with additional predictors to estimate trading volume effectively (Scott and Varian, 2014, 2015). Model 1 included only financial predictors, serving as a baseline. Model 2 incorporated features related to the quantity and quality of WSB submissions, while Model 3 replaced submissions with predictors derived from comments. Finally, Model 4 combined both submissions and comments, allowing us to compare the relative impact of these two discussion formats on trading volume. This approach enabled a direct evaluation of how multimodal inputs—structured market data and unstructured social media discussions—interacted over time, revealing the extent to which online discussions influenced trading behavior.

A key focus of the analysis was assessing whether user submissions or comments provided stronger predictive signals for trading volume fluctuations. The results indicate that submissions, which are typically longer and more structured, exhibited a stronger correlation with trading activity than comments, which are often brief and reactive. Additionally, topic modeling revealed that themes such as "Short Squeeze" and "Hold & Fight" were particularly relevant during periods of increased market volatility, aligning closely with heightened trading activity. These findings suggest that structured textual discussions, rather than fragmented interactions, offer more valuable signals when modeling the link between online discussions and financial markets.

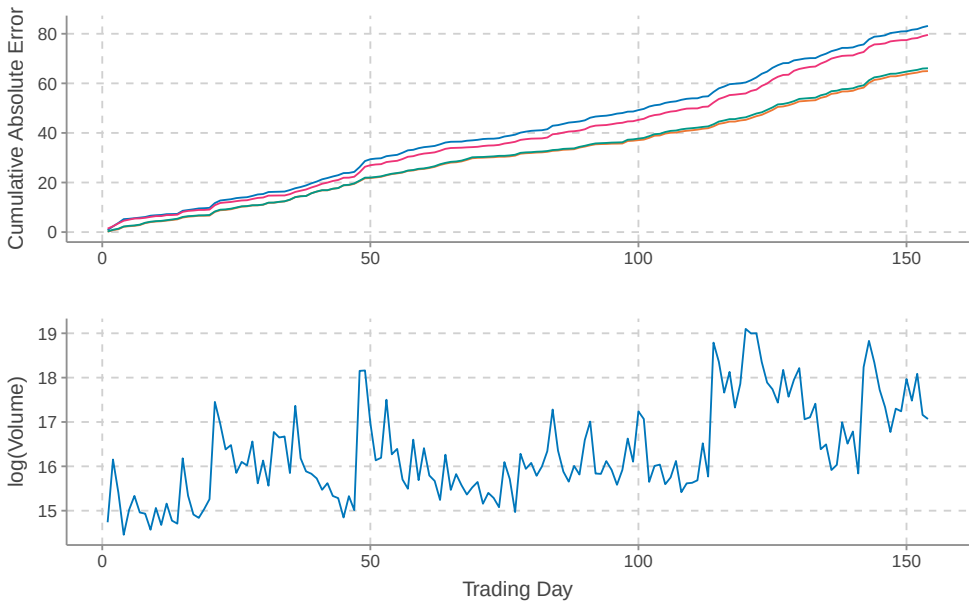

Figure 3.5: Top: Cumulative Absolute Error of Model 1 (blue), Model 2 (orange), Model 3 (red), and Model 4 (green) over Time. Bottom: Trading Volume over Time (Caron et al., 2021, p.11)

As shown in Figure 3.5, incorporating WSB discussions into our BSTS models significantly improved the ability to capture spikes in trading volume. Models that integrated textual features, particularly those derived from submissions, outperformed those relying solely on financial predictors, with the greatest improvements observed during periods of extreme market activity. In particular, Model 2, which incorporated submission-based predictors, outperformed the baseline financial model, improving R^2 (from 0.65 to 0.77) and MAE (from 0.54 to 0.44). These results highlight the importance of multimodal feature representation in financial modeling, demonstrating that integrating structured and unstructured data enhances predictive accuracy. At the same time, these results emphasize the need for specialized preprocessing techniques to encode heterogeneous data sources while maintaining interpretability effectively.

3.5 Paper 4 – Towards a Reliable & Transparent Approach to Data-Driven Brand Valuation

Following the investigation on feature representation in multimodal panel modeling in **Paper 3**, **Paper 4** focuses on addressing similar challenges; however, this time in the context of brand valuation. Like **Paper 3**, this study deals with the complexities of integrating multimodal data – i.e., numerical, categorical, and textual – within a fixed panel structure. However, unlike **Paper 3**, which analyzed data with a high degree of temporal variability, **Paper 4** examines the more constant yearly brand valuations of football clubs in the English Premier League (EPL), providing a structured perspective on temporal trends. Using data from five consecutive seasons, we model brand valuation by operationalizing Aaker's (1991) brand equity framework to capture relevant aspects of brand performance and perception. A key methodological goal of this study is to explore modern feature encoding techniques to enrich the modeling process, demonstrating how proprietary brand valuations can largely be explained and predicted using features from publicly available sources.

To proceed with our analysis, we acquired a dataset spanning five consecutive seasons – i.e., from 2016/2017 to 2020/2021 – of brand valuations for 24 EPL football clubs from Brand Finance. Beyond brand values, the dataset included a wide range of features aligned with Aaker's (1991) brand equity dimensions, such as historical achievements (e.g., total FA Cup wins), squad valuations, social media engagement, and media coverage. These predictors were extracted or generated using structured numerical and

unstructured textual data, capturing the diverse and multimodal nature of brand equity. For instance, social media data captured aspects of *Brand Loyalty*, while textual data from news articles provided insights into *Brand Associations*. Additionally, as shown in Table 3.3, we further enriched the dataset with quantitative features and fan emotions extracted using a Transformer model tailored to Twitter-specific sentiment analysis (Devlin et al., 2019; Liu et al., 2019; Barbieri et al., 2020). Finally, the dataset was structured in a fixed panel format, allowing for a systematic analysis while preserving the temporal dependencies.

Building on this dataset, we employed a mixed-method modeling approach that integrated both explanatory and predictive analyses. For the explanatory analysis, we utilized linear mixed-effects models (also known as hierarchical models) to examine the relationship between brand value and each dimension of brand equity. By incorporating random intercepts for clubs and seasons, these models accounted for the hierarchical structure of the data while capturing sequential dependencies within each club over time. Among the dimensions of brand equity, the *Perceived Quality* model emerged as the most significant, explaining up to 77% of the variation in brand value. Key predictors within this dimension included squad valuation and historical championships, emphasizing the role of on-field success and team composition in shaping brand value. Additionally, while *Brand Awareness* and *Brand Loyalty* contributed less to the overall explanatory power of the model, specific features, such as media coverage and social media engagement, demonstrated significant associations with a valuation trends.

To provide additional insights and to assess practical applications, we employed a gradient boosting model – i.e., XGBoost (Chen and Guestrin, 2016) – for our predictive analysis. This approach utilized all features to estimate brand valuations, achieving robust out-of-sample performance with a mean absolute percentage error (MAPE) of 14%. Interestingly, while the *Perceived Quality* dimension was the most statistically significant in the explanatory analysis, the predictive analysis revealed that features from the *Brand Associations* category, such as club age and stadium age, provided meaningful improvements when combined with other predictors. This contrast highlights a key finding: variables with weaker explanatory power in statistical modeling can still contribute substantially to improving forecasting accuracy when incorporated into predictive frameworks. These results emphasize the importance of considering interpretability and predictive utility when modeling multimodal panel data.

Table 3.3: Descriptive Statistics (Caron et al., 2022, p.6)

		Feature	Measurement	Min.	Median	Mean	Max.	Source	Extracted / ML-Generated
Brand Associations	Club tradition	*Club age*	Years	101.0	133.0	129.2	153.0	Transfermarkt	
		Stadium age	Years	1.0	100.0	78.02	164.0	Transfermarkt	
	Departures (players)	*Player departures*	Total departures p.s.	4.0	17.0	17.5	34.0	Transfermarkt	
		Departures market value	Total market value of departures p.s. / 1,000,000	14.8	90.74	102.50	406.5	Transfermarkt	
	Dismissals (coaches)	*Coach dismissals*	Total dismissal p.s.	0.0	0.0	0.5227	4.0	Transfermarkt	
Brand Awareness	English matches	*EFL Cup matches*	Total matches p.s.	1.0	2.0	2.784	7.0	Transfermarkt	
		FA Cup matches	Total matches p.s.	1.0	3.0	3.227	7.0	Transfermarkt	
	European matches	*Europa League matches*	Total matches p.s.	0.0	0.0	1.5	15.0	Transfermarkt	
		Champions League matches	Total matches p.s.	0.0	0.0	2.33	15.0	Transfermarkt	
	Media coverage	*Guardian articles*	Total articles p.s. / 100	4.130	14.085	15.156	36.470	The Guardian	✓
	Social media presence	*Tweets from club*	Total tweets p.s. / 100	14.34	55.13	54.9	106.85	Twitter	
		Twitter likes from club	Total likes p.s. / 100	0.0	7.14	9.558	55.21	Twitter	
Brand Loyalty	Fan interactions	*Tweets at club*	Total tweets p.s. / 100,000	0.00132	0.42092	1.27991	6.72250	Twitter	
		Ratio of negative tweets at club	Total neg. tweets p.s. / Total tweets p.s.	0.09049	0.15716	0.16826	0.28930	Twitter	✓
		Ratio of positive tweets at club	Total pos. tweets p.s. / Total tweets p.s.	0.2766	0.3637	0.3638	0.4940	Twitter	✓
		Ratio of sensitive tweets at club	Total sens. tweets p.s. / Total tweets p.s.	0.001689	0.005120	0.005520	0.013043	Twitter	✓
	Stadium attendances	*Spectators*	Total spectators p.s. / 10,000	0.0	54.48	56.21	143.05	Twitter	
	Social media followers	*Twitter followers delta*	Total followers Δ p.s. / 10,000	-0.0072	19.8652	79.5535	670.0495	Twitter	
		Reddit subscribers delta	Total subscribers Δ p.s. / 10,000	0.0124	0.28485	1.27948	8.72010	Reddit	
Perceived Quality	Historical English championships	*Hist. Premier League titles*	Total historical championships	0.0	2.0	4.909	20.0	Transfermarkt	
		Hist. FA Cup titles	Total historical championships	0.0	4.0	4.33	14.0	Transfermarkt	
		Hist. EFL Cup titles	Total historical championships	0.0	1.0	2.125	8.0	Transfermarkt	
	Historical European championships	*Hist. Europa League titles*	Total historical championships	0.0	0.0	0.4318	3.0	Transfermarkt	
		Hist. Champions League titles	Total historical championships	0.0	0.0	0.2727	2.0	Transfermarkt	
	Squad valuation	*Squad valuation*	Total market value of squad p.s. / 1,000,000	106.0	307.7	447.0	1203.5	Transfermarkt	
Other Proprietary	Stadium capacity	*Stadium capacity*	Total seats / 10,000	1.133	3.709	4.052	7.488	Transfermarkt	
	Technical sponsor	*Technical sponsor (Adidas)*	Categorical feature		20 / 88 Observations			Misc.	
		Technical sponsor (Nike)	Categorical feature		16 / 88 Observations			Misc.	
		Technical sponsor (Puma)	Categorical feature		18 / 88 Observations			Misc	
		Technical sponsor (Umbro)	Categorical feature		20 / 88 Observations			Misc.	
		*Technical sponsor (Other)**	Categorical feature		14 / 88 Observations			Misc.	

Note: p.s. = per season; * = Reference category

3.6 Paper 5 – Shortcut Learning in Financial Text Mining: Exposing the Overly Optimistic Performance Estimates of Text Classification Models under Distribution Shift

As can be seen from Table 3.1, all works presented in this thesis are likely to be affected by distribution shifts – i.e., systematic changes in the data distribution between training and testing that can undermine a model's generalization capabilities. As exposed in Chapter 2, distribution shifts are particularly concerning in sequential learning problems, where temporal and entity-level dependencies make models especially vulnerable to subtle changes in the underlying data. However, while all prior studies have implicitly faced these issues, they have primarily focused on challenges in feature encoding and data representation, leaving the question of how models handle changing data distributions open.

Hence, in **Paper 5**, we focus on understanding and mitigating the impact of shortcut learning – i.e., a phenomenon where models rely on decision rules that "perform well on [independent and identically distributed (i.i.d.)] test data but fail on [out-of-distribution (o.o.d.)] tests" (Geirhos et al., 2020, p.667). Shortcut learning is inherently tied to distribution shifts, as models exploiting such rules often struggle to generalize when exposed to o.o.d. conditions, revealing the limitations of performance estimates derived solely from i.i.d. evaluations.

To investigate these issues, we focused on financial data, more precisely textual financial data, which provides an ideal testbed given its entity-rich and natural vulnerability to distribution shifts. With dependencies across entities and time being ubiquitous in finance, concepts such as generalization and shortcut learning are especially relevant when developing and evaluating text classification and regression models. Nevertheless, our comprehensive review of ML/NLP-based contributions in financial text mining revealed that concepts like distribution shifts, leakage, or shortcut learning are rarely, if ever, mentioned, much less addressed (Xing et al., 2018; Pejić Bach et al., 2019; Gupta et al., 2020; Mishev et al., 2020). Instead, most studies rely on assumptions of data stability, which fail to account for real-world challenges.

Previous findings have shown that even slight changes in data distribution can significantly degrade model performance, highlighting the critical need for robust evaluation methodologies (Recht et al., 2019; Bastings et al., 2021; Kapoor and Narayanan, 2023). Yet, most ML-based science works rely on naive sampling strategies, such as random

sampling, which draws train and test sets from the same probability distribution (i.i.d.). While such an approach simplifies evaluation, it fails to account for potential distribution shifts and masks critical vulnerabilities in model generalization. Hence, with **Paper 5**, we aim to expose how shortcut learning can undermine the reliability of model evaluations and propose o.o.d. evaluation methodologies for assessing models trained on entity-rich data.

To proceed, we collected and annotated our own dataset of financial microblogs from *Twitter*, addressing the limitations of publicly available datasets, which were often too small in size or lacked the necessary information to generate clear distribution shifts between train and test sets – i.e., a feature defining every document's target entity, or company. Our dataset focused on ten U.S.-based Fortune 500 companies from five distinct sectors, providing a diverse source of textual data.

Figure 3.6: Overview of the Sampling Strategies (Caron, 2022, p.3490)

Building on this dataset, we proposed and tested three sampling strategies to evaluate the impact of distribution shifts on model performance, as illustrated in Figure 3.6. These strategies included random sampling, where train and test sets were drawn randomly from the same distribution; entity-based sampling, which ensures that all documents about a specific company appeared exclusively in either the training or test set, thus highlighting entity-level distribution shifts; and sector-based sampling, which groups companies into industry sectors and ensures that sectors appearing in the training set are distinct from those in the test set. By systematically applying these approaches, we aimed to analyze how different types of distribution shifts affect the generalization capabilities of the various models.

As exemplified in Figure 3.7, the data sampled using the sector-based strategy exhibits clear clustering patterns highlighting substantial differences between subsets. Specifically, the data grouped by sector forms distinct clusters, indicating that posts

related to companies within the same industry share unique linguistic or contextual features. These patterns confirm that the dataset is inherently non-homogeneous, with sector-level factors contributing to its complexity. Such distributional variations visually demonstrate the challenges models face when trained under i.i.d. assumptions and tested on o.o.d. data.

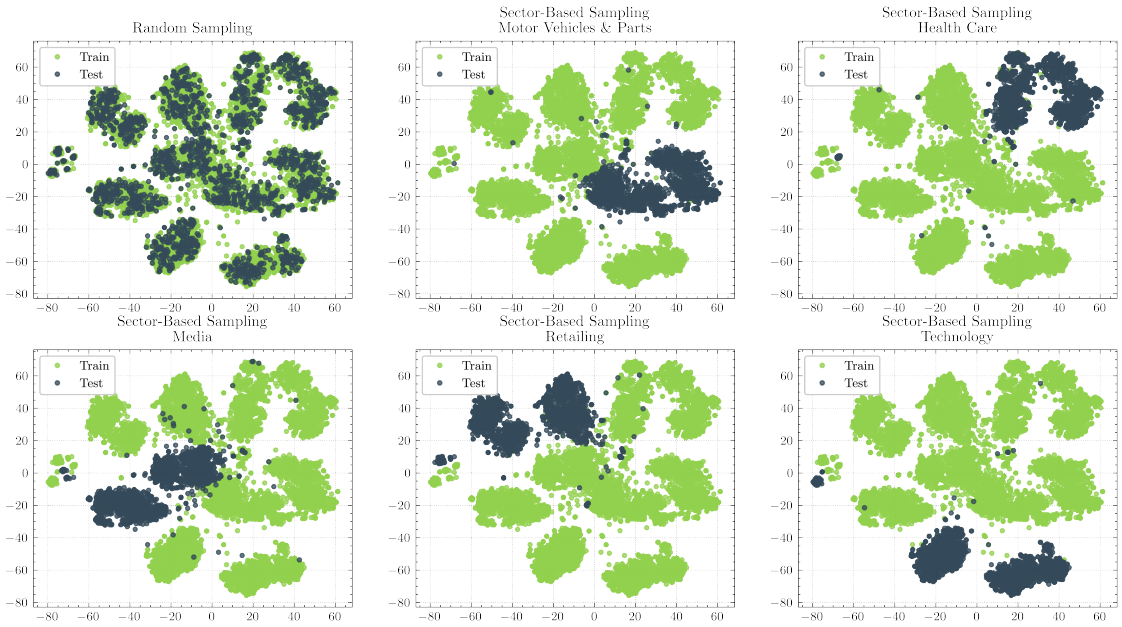

Figure 3.7: t-SNE Visualization of our Out-of-Distribution Tests (o.o.d.) sampled using the Sector-Based Strategy (Caron, 2022, p.3490)

To proceed with our experiments, we applied three distinct preprocessing steps to reduce the likelihood of shortcut learning and provide a more reliable evaluation of model performance under distribution shifts, namely:

- basic preprocessing, which involved removing URLs, unwanted characters, and converting emojis into text to standardize inputs across all models;
- entity removal, where named entities, mentions, and hashtags, such as “Apple” or “Tesla”, were excluded to assess the models’ reliance on entity-specific cues; and
- vocabulary filtering, using a TF-IDF-like approach to eliminate terms strongly associated with specific entities, thereby reducing the influence of entity-specific language patterns.

These preprocessing steps were designed to progressively mitigate shortcut learning and allow for a more thorough assessment of the models’ generalization capabilities.

Finally, to evaluate the impact of the proposed preprocessing techniques and sampling strategies, we fine-tuned four state-of-the-art transformer models – i.e., BERT (Devlin et al., 2019), DistilBERT (Sanh et al., 2019), RoBERTa (Liu et al., 2019), and BART (Lewis et al., 2020). These models were trained with robust optimization techniques, including adaptive learning rates, weight decay regularization, and early stopping, to ensure stability and prevent overfitting. The results revealed a pronounced discrepancy between i.i.d. and o.o.d. performance, with error rates increasing by up to 29.7% under entity-based sampling and 27.2% under sector-based sampling, highlighting the challenges posed by distribution shifts. Preprocessing techniques, such as entity removal and vocabulary filtering, demonstrated their effectiveness in mitigating these effects. Notably, these methods reduced the impact of shortcut learning by as much as 52% and 59% for entity-based and sector-based sampling, respectively. This reduction significantly narrowed the performance gap, providing a more accurate evaluation of the models' true generalization capabilities.

3.7 Paper 6 – Integrating Driver Behavior into Last-Mile Delivery Routing: Combining Machine Learning and Optimization in a Hybrid Decision Support Framework

As outlined in Table 3.1, **Paper 6** explores the challenge of modeling sequential dependencies in spatio-temporal data, where past decisions influence future choices. Capturing these dependencies requires feature encoding techniques that effectively represent evolving decision patterns while preserving the relationships between successive observations. At the same time, output decoding remains a key challenge, as structured predictions need to align with real-world constraints. Hence, this study addresses these complexities by integrating ML-based predictions with optimization-driven prescriptions, demonstrating how sequential patterns can improve structured decision-making.

To investigate these challenges, we developed a hybrid framework that integrates behavioral modeling with decision-support systems, demonstrating how learned sequential patterns can enhance recommendations. Unlike traditional methods that assume decision-makers follow predefined rules, this approach learns from historical sequences, identifying patterns in past choices before generating prescriptions. This is particularly relevant in settings where decision-makers interact repeatedly with an environment, adjusting their behavior based on real-world constraints. For this study, we leveraged the data from the *Amazon Last-Mile Routing Research Challenge (ALMRRC)* (Mer-

chan et al., 2022) – i.e., a large-scale dataset capturing real-world last-mile delivery operations. The dataset provides detailed GPS traces of delivery drivers, allowing us to analyze routing behavior, deviations from optimized paths, and the impact of external constraints such as traffic and parking availability. Rather than assuming full compliance with optimized routes, the framework first predicts the sequence of delivery stops a driver is most likely to follow, then integrates this prediction into an optimization model that refines route recommendations while ensuring behavioral realism.

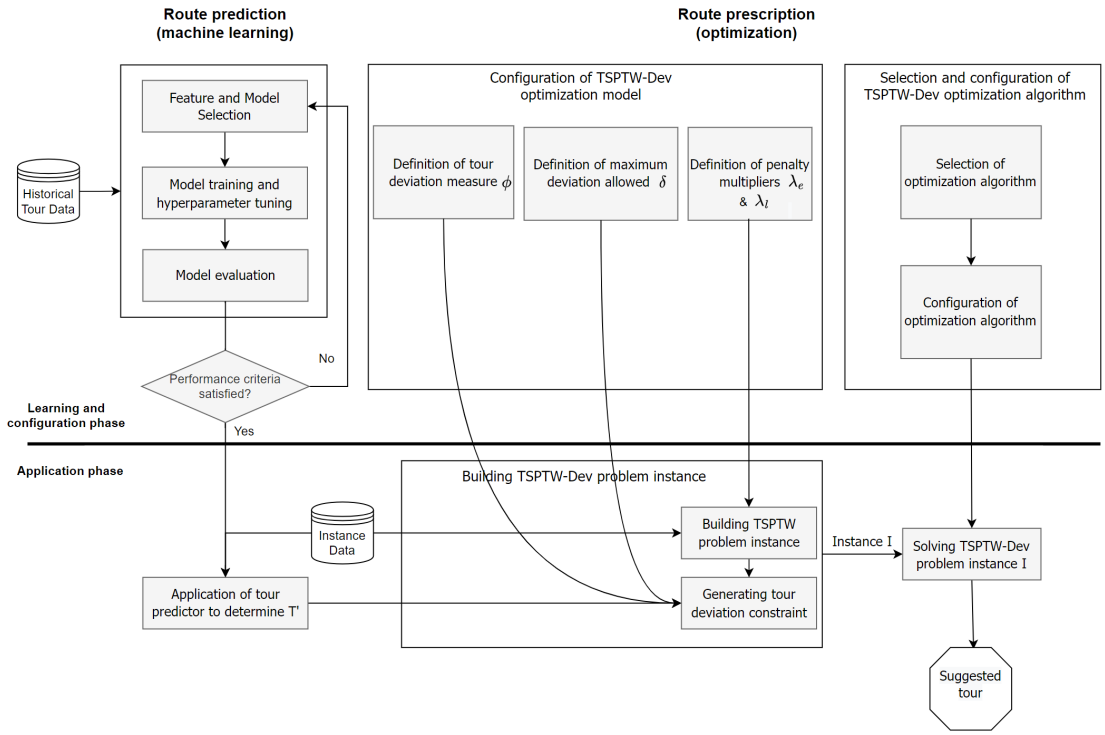

Figure 3.8: Decision Support Framework for Integrating Prediction and Prescription (Dieter et al., 2023, p.288)

Following the framework presented in Figure 3.8, the route prediction step employs ML to forecast the sequence of delivery stops based on historical decision patterns. Each sequence is represented as a series of location pairs enriched with additional contextual features such as travel times, spatial constraints, and individual behavioral tendencies. The prediction model is implemented as a feedforward neural network and trained to rank potential next steps based on learned behavioral patterns. This process is structured in two phases: first, clustering delivery locations into logical service zones, and second, predicting the sequence within each cluster. This hierarchical approach enables the model to generalize across diverse decision-making behaviors while preserving individual routing preferences.

Once the sequence of likely delivery stops is predicted, the optimization step refines these recommendations to balance efficiency with real-world adherence. The predicted sequence is incorporated into a modified Travelling Salesman Problem with Time Windows (TSPTW) model with deviation constraints (Gendreau et al., 1998; Ohlmann and Thomas, 2007; Baldacci et al., 2012) – i.e., *TSPTW-Dev*. This ensures that the suggested sequence remains close to the expected pattern, minimizing the likelihood of non-compliance while optimizing overall efficiency. Deviation constraints, as exemplified in Figure 3.9, are essential to prevent impractical route modifications that could compromise usability, ensuring that suggested solutions remain realistic. To achieve this, Jaro distance and Longest Common Subsequence (LCSS) distance are used as similarity metrics, measuring how closely the optimized solution follows the predicted sequence. The optimization uses a Variable Neighborhood Search (VNS) (Wei et al., 2015) heuristic, which iteratively refines suggestions while enforcing constraints on permitted deviations from the expected behavior.

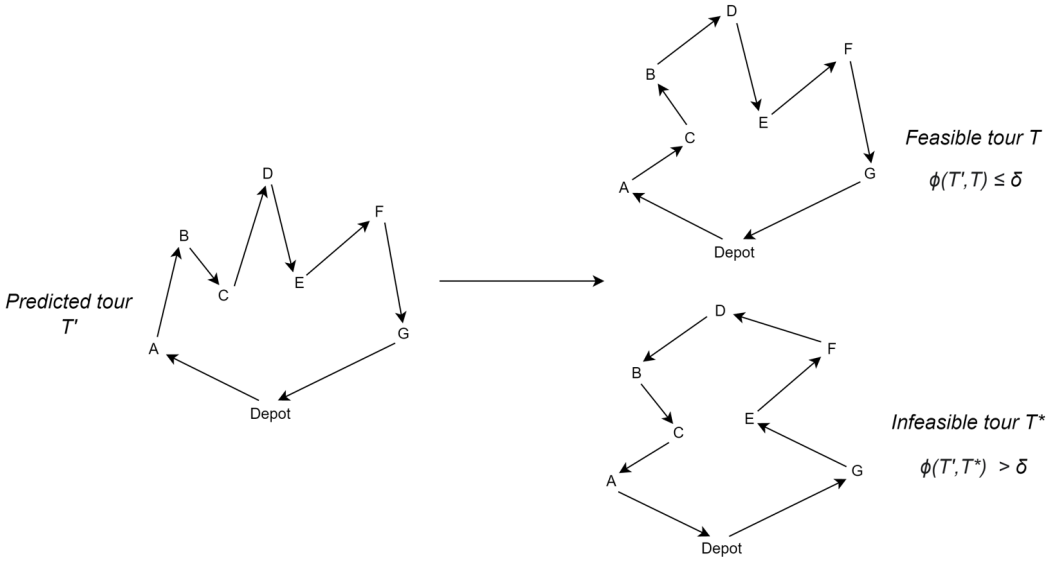

Figure 3.9: Illustrative Example of the Tour Deviation Constraint (Dieter et al., 2023, p.287)

The results in Table 3.4 show that the neural network approach achieves higher accuracy in predicting driver routes than the nearest neighbor baseline. However, finding the right balance in the optimization step is crucial, as allowing more flexibility in route deviations improves efficiency but reduces adherence to expected driver behavior. Our sensitivity analysis on the ALMRRC dataset indicates that even slight deviations from predicted routes can significantly improve travel efficiency. These findings suggest that strict adherence to historical behavior is not always necessary for effective decision-making, aligning with real-world observations where drivers dynamically adjust to traf-

fic, parking constraints, and personal preferences. The study further demonstrates that integrating ML-based route predictions into an optimization framework enhances both predictive accuracy and the quality of suggested routes, reinforcing the value of hybrid decision-support systems in last-mile delivery.

Table 3.4: Tour Prediction Results (Dieter et al., 2023, p.291)

(a) Predicted Tours

	Mean	Median	Std
Jaro	0.306	0.274	0.151
LCSS	0.703	0.716	0.144

(b) Nearest Neighbor Tours

	Mean	Median	Std
Jaro	0.317	0.312	0.126
LCSS	0.729	0.741	0.119

3.8 Paper 7 – TacticalGPT: Uncovering the Potential of LLMs for Predicting Tactical Decisions in Professional Football

As outlined in Table 3.1, **Paper 7** is the only contribution in this thesis that focuses on event-based data – i.e., a format that has become increasingly common in ML for analyzing discrete occurrences over time. Unlike continuous time-series data, where observations follow a fixed temporal structure, event-based sequences unfold dynamically, as discussed in Chapter 2, with irregular spacing and varying levels of contextual importance. This variability presents key challenges in feature encoding, as event logs often lack a standard representation suitable for ML models. Moreover, effectively capturing sequential dependencies is essential, as each event may influence subsequent actions.

To address these issues, this study focuses on how event-based sequences can be structured for machine learning models, particularly in scenarios where multiple actions and outcomes must be predicted from past observations. Football provides an ideal setting for this investigation, as matches unfold through a series of discrete events (e.g., passes, shots, tackles), each carrying implicit dependencies that shape subsequent actions. Consequently, we introduce TacticalGPT – i.e., a fine-tuned Large Language Model (LLM) designed to predict tactical decisions in professional football. By converting structured event data into text-based prompts, we examine how generative approaches can model tactical reasoning through *What-*, *Who-*, and *Where-* type questions, capturing player intent, spatial positioning, and action sequences in a human-interpretable format. At the same time, the model must learn to decode multiple interdependent predictions into structured outputs – i.e., mapping $n \rightarrow m$ relationships – ensuring that responses remain factually consistent and meaningful.

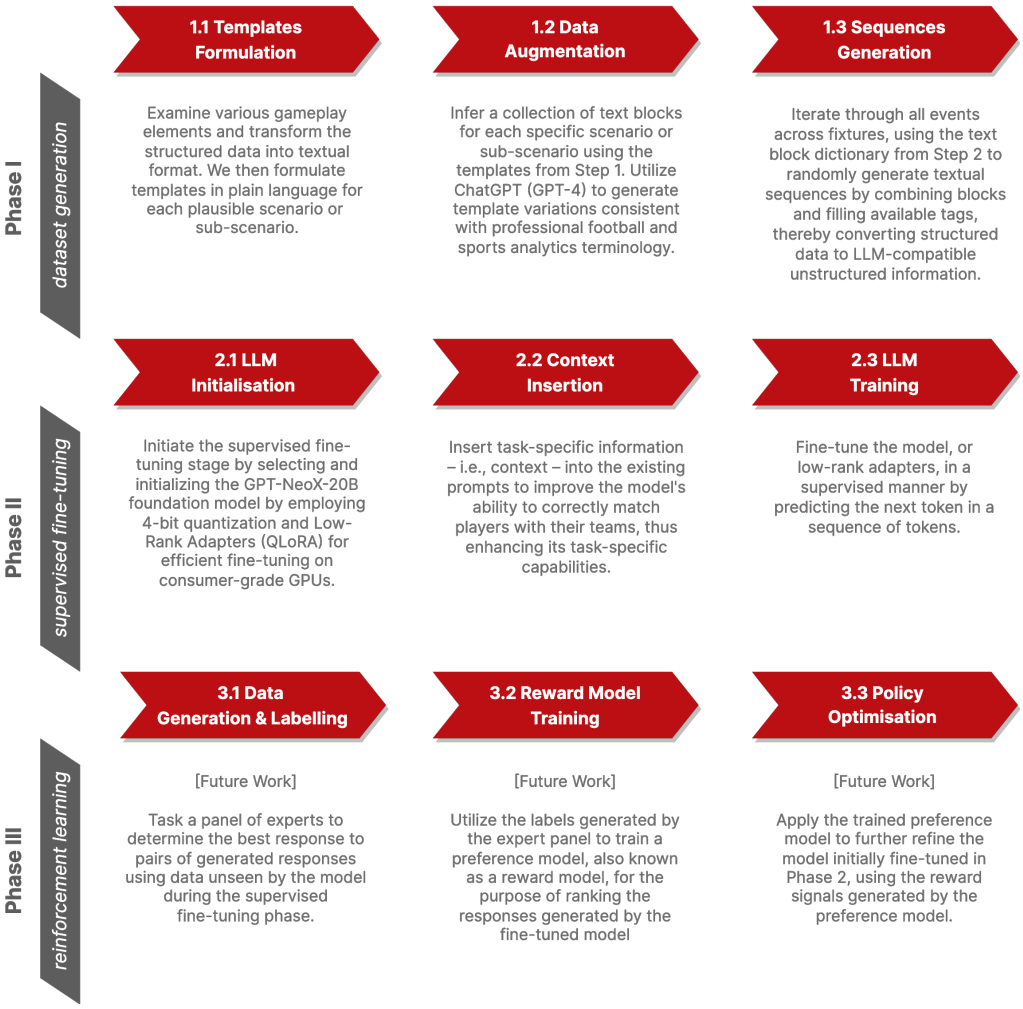

Figure 3.10: TacticalGPT Pipeline (Caron and Müller, 2023, p.4)

Following the pipeline presented in Figure 3.10, the dataset used to fine-tune TacticalGPT was acquired from StatsBomb, covering 580 Premier League matches from the 2021/2022 and 2022/2023 seasons. Each event, such as passes, shots, or tackles, was enriched with metadata, including player identity, event location, and match phase, ensuring that sequential dependencies were preserved. To transform this structured data into a format compatible with natural language models, we used a rule-based system to generate text-based templates that described each event in plain language. These templates incorporated contextual elements, specifying *what* action was performed, *who* was involved, and *where* it occurred on the pitch. The final stage involved sequence generation, where predefined templates were applied to match data to create coherent event sequences. As shown in Figure 3.11, this step ensured that the dataset captured the flow of play while preserving the relationships between successive actions, forming the basis for training TacticalGPT to generate meaningful tactical insights.

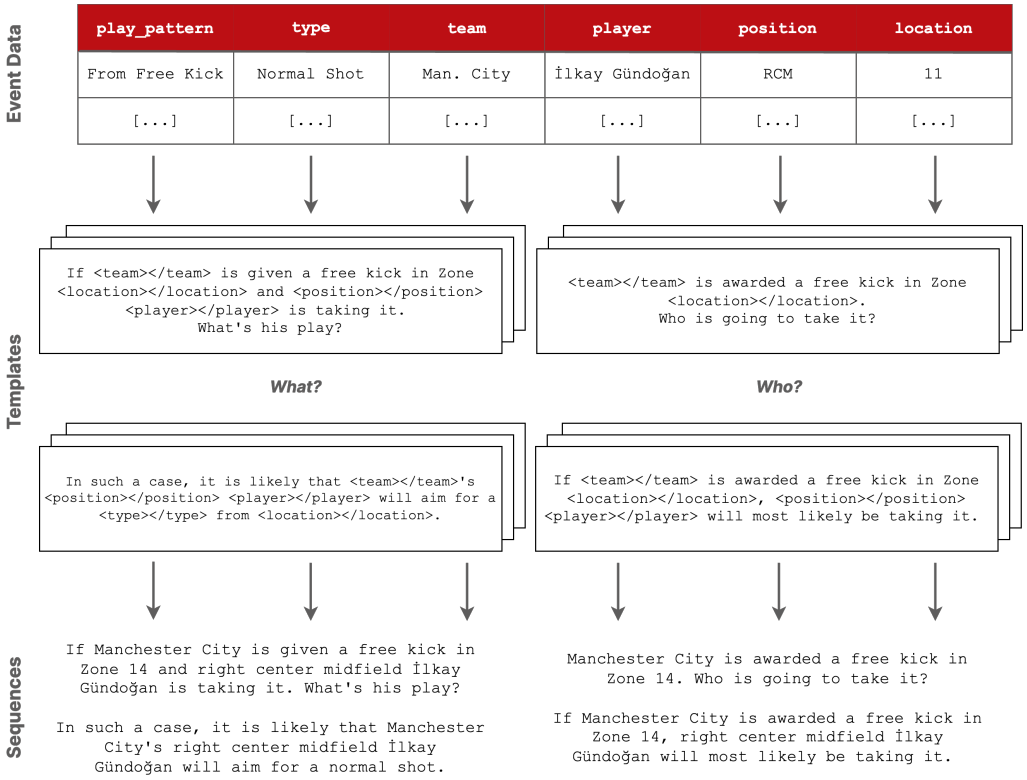

Figure 3.11: Dataset Generation Pipeline for TacticalGPT (Caron and Müller, 2023, p.5)

Moving on to the training process, the model was initialized with the GPT-NeoX-20B foundation model (Black et al., 2022) and fine-tuned using low-rank adapters (QLoRA) – i.e., an efficient fine-tuning method that reduces memory consumption while preserving 16-bit precision for adaptation to specific tasks (Dettmers et al., 2024). To improve the model's understanding of football-specific interactions, information such as team lineups was embedded into prompts, helping TacticalGPT associate players with their teams and actions. The training followed a next-token prediction objective, enabling the model to generate fluent and tactically coherent responses from event-based sequences. To enhance robustness, TacticalGPT was fine-tuned on 100,000 artificial event sequences, incorporating synthetic variations to align its predictions with real-world football decision-making. Finally, adaptive learning rates and early stopping were applied to optimize performance and prevent overfitting, ensuring that the model generalizes well to unseen match scenarios.

With the supervised training complete, the final stage of the pipeline turns to refining TacticalGPT's ability to generate high-quality, contextually accurate outputs. Phase 3, centered on output decoding, introduces reinforcement learning as a critical step for aligning the model's predictions with real-world scenarios. While this phase rep-

resents future work, it aims to incorporate expert-labeled responses to train a reward model, enabling TacticalGPT to produce more precise and reliable outputs. This step addresses challenges like spatial predictions, where accuracy and contextual alignment are essential for actionable insights.

TacticalGPT demonstrates strong performance in generating responses for *What*- and *Who*-type questions, with a high rate of factually correct and factually plausible answers. Specifically, for *What*-type questions, 50% of responses were factually correct – i.e., fully matching the ground truth – while 46% were factually plausible, meaning they differed slightly but remained reasonable within the context. Similarly, for *Who*-type questions, 32% were factually correct, and 62% were factually plausible. However, a small portion of responses – i.e., 4% for *What* and 6% for *Who* – were factually improbable, significantly diverging from the ground truth in ways that were unlikely to be true. However, *Where*-type questions proved more challenging, as the model struggled with precise spatial predictions. While it effectively captured player actions and tactical decisions, it had difficulties determining exact event locations. These results highlight the model's ability to process sequential data while revealing challenges in encoding and decoding positional information.

3.9 Paper 8 – Detecting and Mitigating Shortcut Learning Bias in Machine Learning: A Pathway to More Generalizable ML-based (IS) Research

As introduced in **Paper 5**, shortcut learning and distribution shifts represent fundamental challenges in machine learning, often remaining undetected and affecting models across all domains. These issues are not only problematic in laboratory conditions but also lead to failures in real-world deployments. While modern ML models frequently achieve high accuracy on benchmark datasets, they tend to exploit spurious correlations rather than learning true causal relationships, resulting in significant performance degradation under changing conditions. As outlined in Table 3.1, these challenges are present across multiple contributions in this thesis, as well as in broader machine learning research. For instance, panel data settings in **Paper 1**, **Paper 3**, and **Paper 4** may expose models to the risk of overfitting entity-specific patterns or temporal dependencies, limiting their generalization across unseen instances. Similarly, **Paper 5**, which focuses exclusively on textual data, illustrates how transformer-based models trained on entity-rich financial text often rely on superficial cues rather than meaningful semantic

patterns, leading to misleadingly high accuracy under standard evaluations. **Paper 8** extends these discussions by introducing a systematic framework to detect and mitigate shortcut learning in tabular data, a setting that has received far less attention despite being a common ML application in fields such as finance, healthcare, and IS.

To address these challenges, we developed, in this study, a two-phase framework that helps detect (Phase 1) and mitigate (Phase 2) shortcut learning in machine learning models. Unlike prior work focusing on computer vision and NLP, this framework is designed explicitly for tabular data, which remains one of the most widely used data types in applied machine learning. As shown in Figure 3.12, Phase 1 systematically evaluates model performance under different distributional settings – i.e., i.i.d., time-based, entity-based, and combined time/entity-based sampling – to assess the extent of shortcut dependency. This phase also emphasizes transparent reporting, ensuring that model evaluations expose shortcut dependencies rather than concealing them under standard overoptimistic i.i.d. assumptions. Phase 2, which is optional and iterative, introduces a viable feature exclusion strategy as a demonstration, leveraging the Classifier Two-Sample Test (C2ST) to identify and remove shortcut-prone features. Since mitigation can be repeated multiple times, the framework allows for flexibility in adjusting feature representations while maintaining proper reporting at every stage, regardless of whether mitigation is applied.

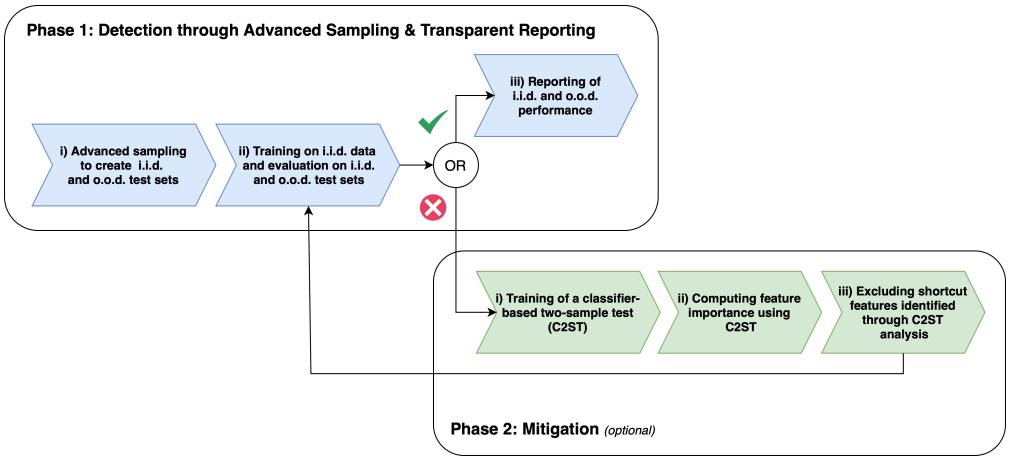

Figure 3.12: A Structured Approach to Detecting, Mitigating, and Reporting Shortcut Learning in ML-Based Research (Caron et al., 2025, p.7)

In addition to testing the proposed framework on simulated data, we applied it to the real-world classification task of corporate credit risk prediction – i.e., a problem widely studied in the IS, operations research (OR), and finance literature. Our dataset consisted of 12,533 observations from 1,702 unique firms spanning over 16 years

(2000–2016), helping us capture firm-level financial indicators commonly used in credit evaluations. Five widely used machine learning models, namely Logistic Regression, Random Forest, XGBoost, TabNet, and AutoML, were used to analyze how different models respond to shortcut learning.

To systematically identify shortcut dependencies, the framework employs a structured evaluation process based on advanced sampling strategies (as shown in Figure 3.13). Following a five-fold cross-validation approach, the dataset was segmented into different training and test splits – i.e., i.i.d., o.o.d. (time-based), o.o.d. (entity-based), and o.o.d. (time/entity-based) – to examine how predictive performance varies under distributional shifts. To ensure fair comparisons across these sampling strategies, Gaussian-based Bayesian optimization is used to tune model hyperparameters separately for each fold. By systematically assessing model performance across these different conditions, this phase exposes how much models rely on spurious associations, uncovering shortcut dependencies that would otherwise remain hidden under conventional i.i.d. evaluations.

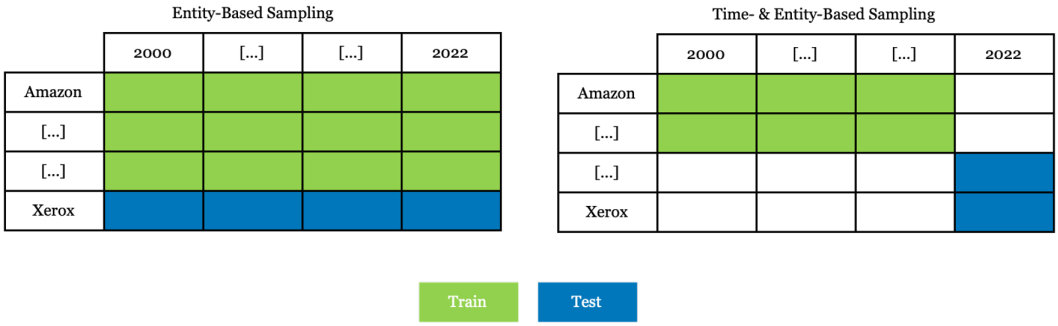

Figure 3.13: Advanced Sampling Strategies (Caron et al., 2025, p.9)

As an optional next step, Phase 2 (Mitigation) focuses on a feature exclusion approach to reduce shortcut dependencies. This process demonstrates one possible mitigation strategy but is not a definitive solution, as different applications may require alternative approaches. The Classifier Two-Sample Test (C2ST), initially designed for evaluating Generative Adversarial Networks (GANs), is used to detect features exhibiting significant distribution shifts between i.i.d. and o.o.d. datasets. Once identified, shortcut-prone features can be removed, allowing models to learn more generalizable patterns rather than relying on dataset artifacts. Since mitigation can be repeated iteratively, it offers flexibility in refining model robustness, ensuring that improvements in generalization do not come at the cost of excessive information loss. Regardless of whether mitigation is applied, the framework mandates transparent reporting across all sampling strategies, preventing inflated performance estimates from i.i.d.-biased evaluations.

Table 3.5: Predictive performance of different models on the corporate credit rating data for various i.i.d. and o.o.d. test sets (Caron et al., 2025, p.18)

	Sampling strategy	**Predictive performance**			**Relative change vs. i.i.d. (test)**		
		$Error_{bal.}$	$F1_{macro}$	$G\text{-}Mean_{macro}$	$Error_{bal.}$	$F1_{macro}$	$G\text{-}Mean_{macro}$
Logistic Regression	i.i.d. (test)	0.3999	0.4509	0.7305	—	—	—
	o.o.d. (time)	0.4379	0.4202	0.6971	-9.07%	-7.04%	-4.68%
	o.o.d. (entity)	0.5115	0.3691	0.647	-24.49%	-19.95%	-12.12%
	o.o.d. (time/entity)	0.5489	0.3159	0.6017	-31.41%	-35.21%	-19.34%
Random Forest	i.i.d. (test)	0.3545	0.5574	0.7549	—	—	—
	o.o.d. (time)	0.3719	0.5061	0.7424	-4.79%	-9.65%	-1.67%
	o.o.d. (entity)	0.4424	0.4588	0.6979	-22.06%	-19.41%	-7.85%
	o.o.d. (time/entity)	0.5073	0.4186	0.6494	-35.46%	-28.44%	-15.03%
XGBoost	i.i.d. (test)	0.3131	0.6396	0.7849	—	—	—
	o.o.d. (time)	0.3358	0.6113	0.7691	-7.00%	-4.52%	-2.03%
	o.o.d. (entity)	0.5078	0.4531	0.6635	-47.44%	-34.14%	-16.76%
	o.o.d. (time/entity)	0.5579	0.4192	0.6244	-56.21%	-41.63%	-22.78%
TabNet	i.i.d. (test)	0.3640	0.5916	0.7590	—	—	—
	o.o.d. (time)	0.3655	0.5773	0.7562	-0.41%	-2.45%	-0.37%
	o.o.d. (entity)	0.5624	0.3994	0.6153	-42.83%	-38.79%	-20.91%
	o.o.d. (time/entity)	0.5757	0.3711	0.6046	-45.06%	-45.81%	-22.65%
AutoML	i.i.d. (test)	0.2899	0.7142	0.8217	—	—	—
	o.o.d. (time)	0.3563	0.6487	0.7785	-20.55%	-9.61%	-5.4%
	o.o.d. (entity)	0.5933	0.4034	0.5998	-68.7%	-55.62%	-31.22%
	o.o.d. (time/entity)	0.5971	0.3932	0.5897	-69.27%	-57.97%	-32.88%

As shown in Table 3.5, performance degradation varied significantly across different models and sampling strategies. For instance, under i.i.d. evaluation, models such as AutoML and XGBoost exhibited strong predictive performance, with F1-macro scores of 0.713 and 0.639, respectively. However, when tested using o.o.d. (time/entity) samples, performance dropped by as much as 57.97% for AutoML and 41.63% for XGBoost,

illustrating their reliance on spurious correlations. This pattern was consistent across all evaluated models, confirming that shortcut learning leads to severe generalization failures when the data distribution changes. Interestingly, time-based distribution shifts alone had a relatively moderate impact, with models such as Random Forest and TabNet experiencing performance declines of less than 10%. In contrast, entity-based and combined time/entity-based shifts caused significantly more significant degradation. These findings underscore the necessity of evaluating models under diverse distributional conditions, as conventional i.i.d. assessments may mask vulnerabilities that become evident only in o.o.d. scenarios.

4 Discussion & Conclusion

4.1 Implications for Research & Practice

From a methodological perspective, this thesis contributes to sequential data modeling by tackling key challenges related to feature encoding, output decoding, and distribution shifts. As exposed in the previous chapters, our results show that how models process sequential data plays a key role in improving predictive accuracy and adaptability across different applications. In practice, ensuring that machine learning models remain effective outside controlled settings requires methods that process data efficiently, produce structured predictions, and account for changes in data over time.

Consequently, an important part of this research is the study of feature encoding – i.e., how models represent input data. Our results show that encoding decisions impact performance, as models make better predictions when they preserve relationships between observations instead of using fixed data representations. Hence, this thesis examined how encoding methods can be adapted to different data types, such as long-text documents or spatio-temporal sequences. For instance, **Paper 1** demonstrates how integrating textual information into financial models improves predictive accuracy, while **Paper 2** highlights the benefits of learning movement patterns directly from tracking data rather than relying on predefined event labels. These findings show that selecting an encoding approach that matches the structure of the data helps models learn from sequential information and capture dependencies between observations more effectively. For practitioners, this means that encoding choices should be carefully considered when integrating diverse data sources in real-world applications.

Moreover, this thesis also contributes to the understanding of output decoding – i.e., how models produce predictions that follow the structure and constraints of the problem. Many tasks, such as spatio-temporal forecasting or event sequence modeling, require structured outputs. This research shows that decoding methods influence how well models generate predictions that align with real-world patterns. For example, **Paper 7** examines how event-based models transform irregular sequences into structured outputs, demonstrating how decoding strategies affect how models capture event and

temporal dependencies. By comparing different approaches, this work helps identify strategies to improve model reliability and effectiveness for sequential data. In practical applications, ensuring that models generate meaningful and interpretable outputs is critical, particularly in domains where predictions directly impact decision-making.

Lastly, this thesis also examines the challenge of distribution shifts – i.e., where models trained on one dataset struggle when the data distribution changes over time or across entities. As discussed, most standard evaluation methods assume that training and test data follow the same distribution, which, as shown earlier, can lead to overly optimistic performance estimates. This research shows how predictive accuracy can drop in o.o.d. settings due to temporal or entity shifts in the data, as demonstrated in **Paper 5** and **Paper 8**. To address this, we proposed a framework for detecting, mitigating, and reporting shortcut learning, ensuring that models are evaluated under conditions that reveal their strengths and weaknesses rather than relying on artificially inflated results. This structured approach is equally relevant to researchers and practitioners since failing to account for distribution shifts when deploying machine learning models in dynamic environments, such as finance, healthcare, or logistics, can lead to costly or life-threatening errors. By making these risks explicit and providing a structured approach to handling them, this work contributes to more reliable and responsible model deployment in real-world applications.

Together, these contributions add to the literature on sequential data modeling by showing that encoding choices affect what a model learns, decoding methods impact how well predictions fit real-world requirements, and model evaluation must account for distribution shifts to avoid misleading results. Addressing these challenges helps build reliable models beyond controlled experiments and ensures that machine learning systems can be trusted in practical applications where accuracy and robustness matter.

4.2 Limitations

While this thesis advances sequential data modeling across multiple domains, some limitations remain. First, although the research covers a range of data types, it does not address all challenges in sequential learning. Certain aspects, such as long-term dependencies in highly irregular sequences, require further investigation.

Second, while this thesis introduces effective feature encoding strategies and evaluation methodologies, it is based on specific datasets and modeling assumptions. For instance,

the proposed framework and empirical findings depend on the characteristics of the employed datasets, which may not fully generalize to other applications. For example, the event-based modeling approach in **Paper 7** shows strong results in football analytics but may need adjustments for other sports or domains with different event structures.

Finally, the research on distribution shifts – i.e., **Paper 5** and **Paper 8** – mainly focuses on tabular and textual data, providing fewer insights into how these challenges affect spatio-temporal and event data. While the evaluation framework can help detect and mitigate shortcut learning, it does not offer a universal solution. Other techniques could, in fact, further improve model robustness.

4.3 Future Directions & Outlook

Although this thesis addresses key challenges in sequential data modeling, two main areas require further exploration. The first is improving feature encoding strategies, particularly for highly irregular sequences and multimodal dependencies. While this work provides practical solutions for different types of sequential data, handling more complex relationships, such as hierarchical structures or graph-based dependencies, remains an open challenge. Hence, developing encoding methods to capture such relationships better could enhance model performance across various applications.

The second area concerns model generalization under distribution shifts, which, as exposed earlier, remains a crucial issue in ML. While this thesis introduces an evaluation framework to detect and mitigate shortcut learning, further research is needed to refine these approaches and extend them to additional data types. In particular, understanding how distribution shifts affect spatio-temporal and event data is critical, as patterns may evolve differently compared to tabular or textual data. A broader analysis of these shifts could help improve evaluation practices and model reliability in dynamic environments where data distributions constantly change.

To conclude, we firmly believe that advancing these areas will further strengthen sequential data modeling, ensuring that ML models remain effective in research and practical applications. The ability to process sequential data accurately, generate structured predictions, and adapt to changing conditions is essential for reliable performance across different use cases. By building on the contributions of this thesis, future research can continue refining these methods, making ML models more adaptable, reliable, and better suited for complex real-world scenarios.

Part B

Research Papers

Hardening Soft Information: A Transformer-Based Approach to Forecasting Stock Return Volatility

Matthew Caron
Paderborn University

Oliver Müller
Paderborn University

Abstract—Historically, the field of financial forecasting almost exclusively relied on so-called hard information – i.e., numerical data with well-defined and unambiguous meaning. Over the last few decades, however, researchers and practitioners alike have, following the advances in natural language understanding, started recognizing the benefits of integrating soft information into financial modelling. In line with the above, this paper examines whether contemporary attention-based sequence-to-sequence models, known as Transformers, can help improve stock return volatility prediction when applied to corporate annual reports. Using a publicly available benchmark dataset, we show, in an empirical analysis, that out-of-the-box Transformer models have the ability to outmatch current state-of-the-art results and, more importantly, that our proposed feature-based Transformer approach can outperform a robust numerical baseline. To the best of our knowledge, this is the first empirical study focusing on stock return volatility prediction (1) to ever experiment with state-of-the-art Transformer architectures and (2) to demonstrate that a model based solely on soft information can surpass its numerical counterpart. Furthermore, we show that by including an additional numerical feature into our best text-only model, we can push the performance of our model even further, suggesting that soft and hard information contain different predictive signals.

Full Citation: *Caron, M. and Müller, O.(2020). Hardening Soft Information: A Transformer-Based Approach to Forecasting Stock Return Volatility. In Proceedings of the IEEE International Conference on Big Data, pages 4383–4391.*

PIVOT: A Parsimonious End-to-End Learning Framework for Valuing Player Actions in Handball using Tracking Data

Matthew Caron
Paderborn University

Oliver Müller
Paderborn University

Michael Döring
SG Flensburg-Handewitt

Tim Heuwinkel
Paderborn University

Jochen Baumeister
Paderborn University

Abstract—Over the last years, several approaches for the data-driven estimation of expected possession value (EPV) in basketball and association football (soccer) have been proposed. In this paper, we develop and evaluate PIVOT: the first such framework for team handball. Accounting for the fast-paced, dynamic nature and relative data scarcity of handball, we propose a parsimonious end-to-end deep learning architecture that relies solely on tracking data. This efficient approach is capable of predicting the probability that a team will score within the near future given the fine-grained spatio-temporal distribution of all players and the ball over the last seconds of the game. Our experiments indicate that PIVOT is able to produce accurate and calibrated probability estimates, even when trained on a relatively small dataset. We also showcase two interactive applications of PIVOT for valuing actual and counterfactual player decisions and actions in real-time.

Full Citation: *Müller, O., Caron, M., Döring, M., Heuwinkel, T., and Baumeister, J. (2021). PIVOT: A Parsimonious End-to-End Learning Framework for Valuing Player Actions in Handball using Tracking Data. In Proceedings of Workshop on Machine Learning and Data Mining for Sports Analytics (ECML PKDD), pages 116–128.*

To the Moon! Analyzing the Community of "Degenerates" Engaged in the Surge of the GME Stock

Matthew Caron
Paderborn University

Maryna Gulenko
Paderborn University

Oliver Müller
Paderborn University

Abstract—In early 2021, the finance world was taken by storm by the dramatic price surge of the GameStop Corp. stock. This rise is being, at least in part, attributed to a group of Redditors belonging to the now-famous *r/wallstreetbets* (WSB) subreddit group. In this work, we set out to address if user activity on the WSB subreddit is associated with the trading volume of the GME stock. Leveraging a unique dataset containing more than 4.9 million WSB posts and comments, we assert that user activity is associated with the trading volume of the GameStop stock. We further show that posts have a higher predictive power than comments and are especially helpful for predicting unusually high trading volume. Lastly, as recent events have shown, we believe that these findings have implications for retail and institutional investors, trading platforms, and policymakers, as these can have disruptive potential.

Full Citation: *Caron, M., Gulenko, M., and Müller, O. (2021). To the Moon! Analyzing the Community of "Degenerates" Engaged in the Surge of the GME Stock. In Proceedings of the International Conference on Information Systems, pages 2432–2448.*

Towards a Reliable & Transparent Approach to Data-Driven Brand Valuation

Matthew Caron
Paderborn University

Christian Bartelheimer
Paderborn University

Oliver Müller
Paderborn University

Abstract—Now accounting for more than 80% of a firm's worth, brands have become essential assets for modern organizations. However, methods and techniques for the monetary valuation of brands are still under-researched. Hence, the objective of this study is to evaluate the utility of explanatory statistical models and machine learning approaches for explaining and predicting brand value. Drawing upon the case of the most valuable English football brands during the 2016/17 to 2020/21 seasons, we demonstrate how to operationalize Aaker's (1991) theoretical brand equity framework to collect meaningful qualitative and quantitative feature sets. Our explanatory models can explain up to 77% of the variation in brand valuations across all clubs and seasons, while our predictive approach can predict out-of-sample observations with a mean absolute percentage error (MAPE) of 14%. Future research can build upon our results to develop domain-specific brand valuation methods while enabling managers to make better-informed investment decisions.

Full Citation: *Caron, M., Bartelheimer, C., and Müller, O. (2022). Towards a Reliable & Transparent Approach to Data-Driven Brand Valuation. In Proceedings of the Americas Conference on Information Systems, pages 1353–1363.*

Shortcut Learning in Financial Text Mining: Exposing the Overly Optimistic Performance Estimates of Text Classification Models under Distribution Shift

Matthew Caron
Paderborn University

Abstract—In recent years, many cases of deep neural networks failing dramatically when faced with adversarial or real-world examples have been reported. Such failures, which are quite hard to detect, are often related to a generalization problem known as shortcut learning. Yet, with state-of-the-art transformer models now being ubiquitous in financial text mining, one cannot help but wonder how reliable the results conveyed in the ever-growing literature genuinely are. Against this background, we expose, in this work, how vulnerable contemporary financial text mining approaches are to shortcut learning. Focussing on the common learning task of financial sentiment classification, we assess, using two entity-based sampling strategies and our publicly-available dataset, the discrepancies between i.i.d. and o.o.d. performance estimates of four transformer models. Our results reveal that o.o.d. performance estimates are consistently weaker than those of their i.i.d. counterparts, with the error rate increasing by as much as 29.7%, thus, demonstrating how this issue can, when overlooked, lead to misleading evaluations. Moreover, we show how additional preprocessing steps, such as entity removal and vocabulary filtering, can help reduce the effects of shortcut learning by filtering out entity-related linguistic cues.

Full Citation: *Caron, M. (2022). Shortcut Learning in Financial Text Mining: Exposing the Overly Optimistic Performance Estimates of Text Classification Models under Distribution Shift. In Proceedings of the IEEE International Conference on Big Data, pages 3486–3495.*

Integrating Driver Behavior into Last-Mile Delivery Routing: Combining Machine Learning and Optimization in a Hybrid Decision Support Framework

Peter Dieter
Paderborn University

Matthew Caron
Paderborn University

Guido Schryen
Paderborn University

Abstract—The overall quality of last-mile delivery in terms of operational costs and customer satisfaction is primarily affected by traditional logistics planning and the consideration and integration of driver knowledge and behavior. However, this integration has yet to be exploited. This phenomenon is mirrored in two largely separated research bodies on logistics planning and driver behavior. Bridging this gap by using and integrating historical data from actually driven tours into last-mile delivery planning is promising for research and practice. Still, it also leads to complex and large-scale routing problems, which require the development of an overall methodology that goes beyond classical optimization approaches as the needed approach requires a multi-stakeholder perspective, calls for a hybrid-analytical approach by incorporating tour prediction and prescription, and requires both data science and optimization methods. Accounting for these challenges, we suggest a hybrid decision support framework for the traveling salesman problem with time windows that combines machine learning techniques and conventional optimization methods and considers the deviation between suggested and predicted tours. We demonstrate the applicability of our framework in a case study that draws on real-world logistics data. Relying on a sensitivity analysis, we investigate and illustrate the trade-off between the level of deviation between predicted and suggested tours and tour costs. Our case study draws general managerial implications and recommendations that guide decision makers in building their decision support systems for last-mile delivery routing by instantiating our generic framework.

Full Citation: *Dieter, P., Caron, M., and Schryen, G. (2023). Integrating Driver Behavior into Last-Mile Delivery Routing: Combining Machine Learning and Optimization in a Hybrid Decision Support Framework. European Journal of Operational Research, 311(1).*

TacticalGPT: Uncovering the Potential of LLMs for Predicting Tactical Decisions in Professional Football

Matthew Caron
Paderborn University

Oliver Müller
Paderborn University

Abstract—N/A

Full Citation: *Caron, M. and Müller, O. (2023). TacticalGPT: Uncovering the Potential of LLMs for Predicting Tactical Decisions in Professional Football. In Proceedings of the StatsBomb Conference.*

Detecting and Mitigating Shortcut Learning Bias in Machine Learning: A Pathway to More Generalizable ML-based (IS) Research

Matthew Caron
Paderborn University

Oliver Müller
Paderborn University

Johannes Kriebel
University of Hamburg

Abstract—Shortcut learning is a critical challenge in machine learning (ML) that arises when models rely on spurious patterns or superficial associations rather than meaningful relationships in the data. While this issue has been widely studied in computer vision and natural language processing, its impact on tabular and categorical data – i.e., data common in ML-based research within Information Systems (IS) – remains underexplored. To address this challenge, we propose a two-phase framework: detecting shortcut learning biases through advanced sampling strategies and mitigating these biases using methods like feature exclusion. Additionally, we emphasize the importance of transparent reporting to enhance reproducibility and provide insights into a model's generalization capabilities. Using simulated and real-world data, we demonstrate the harmful effects of shortcut learning in tabular data. The results highlight how distribution shifts expose shortcut dependencies, a key focus of the detection phase in our framework. These shifts reveal how models relying on shortcuts fail to generalize beyond training data. While our mitigation strategy is exploratory, it demonstrates that addressing shortcut learning is feasible and underscores the need for further research into model-agnostic solutions. By encouraging comprehensive evaluations and transparent reporting, this work aims to advance the generalizability, reproducibility, and reliability of ML-based research in IS.

Full Citation: *Caron, M., Müller, O., and Kriebel, J. (2025). Detecting and Mitigating Shortcut Learning Bias in Machine Learning: A Pathway to More Generalizable ML-based (IS Research. Working Paper Series, Paderborn University, Faculty of Business Administration and Economics, (129).*

Bibliography

Aaker, D. A. (1991). *Managing Brand Equity*. Free Press.

Anzer, G. and Bauer, P. (2021). A goal scoring probability model for shots based on synchronized positional and event data in football (soccer). *Frontiers in Sports and Active Living*, 3:624475.

Atluri, G., Karpatne, A., and Kumar, V. (2018). Spatio-Temporal Data Mining: A Survey of Problems and Methods. *ACM Computing Surveys (CSUR)*, 51(4):1–41.

Baldacci, R., Mingozzi, A., and Roberti, R. (2012). New State-Space Relaxations for Solving the Traveling Salesman Problem with Time Windows. *INFORMS Journal on Computing*, 24(3):356–371.

Barbieri, F., Camacho-Collados, J., Neves, L., and Espinosa-Anke, L. (2020). TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification. *arXiv preprint arXiv:2010.12421*.

Bastings, J., Ebert, S., Zablotskaia, P., Sandholm, A., and Filippova, K. (2021). Will You Find These Shortcuts? A Protocol for Evaluating the Faithfulness of Input Salience Methods for Text Classification. *arXiv preprint arXiv:2111.07367*.

Beltagy, I., Peters, M. E., and Cohan, A. (2020). Longformer: The Long-Document Transformer. *arXiv preprint arXiv:2004.05150*.

Black, S., Biderman, S., Hallahan, E., Anthony, Q., Gao, L., Golding, L., He, H., Leahy, C., McDonell, K., Phang, J., Pieler, M., Prashanth, U. S., Purohit, S., Reynolds, L., Tow, J., Wang, B., and Weinbach, S. (2022). GPT-NeoX-20B: An Open-Source Autoregressive Language Model. In Fan, A., Ilic, S., Wolf, T., and Gallé, M., editors, *Proceedings of the Workshop on Challenges & Perspectives in Creating Large Language Models*, pages 95–136.

Box, G. E. P., Jenkins, G. M., and Reinsel, G. C. (2008). *Time Series Analysis*. John Wiley & Sons, Ltd.

Caron, M. (2022). Shortcut Learning in Financial Text Mining: Exposing the Overly Optimistic Performance Estimates of Text Classification Models under Distribution Shift. In *Proceedings of the IEEE International Conference on Big Data*, pages 3486–3495.

Caron, M., Bartelheimer, C., and Müller, O. (2022). Towards a Reliable & Transparent Approach to Data-Driven Brand Valuation. In *Proceedings of the Americas Conference on Information Systems*, pages 1353–1363.

Caron, M., Gulenko, M., and Müller, O. (2021). To the Moon! Analyzing the Community of "Degenerates" Engaged in the Surge of the GME Stock. In *Proceedings of the International Conference on Information Systems*, pages 2432–2448.

Caron, M. and Müller, O. (2020). Hardening Soft Information: A Transformer-Based Approach to Forecasting Stock Return Volatility. In *Proceedings of the IEEE International Conference on Big Data*, pages 4383–4391.

Caron, M. and Müller, O. (2023). TacticalGPT: Uncovering the Potential of LLMs for Predicting Tactical Decisions in Professional Football. In *Proceedings of the StatsBomb Conference*.

Caron, M., Müller, O., and Kriebel, J. (2025). Detecting and Mitigating Shortcut Learning Bias in Machine Learning: A Pathway to More Generalizable ML-based (IS) Research. *Working Paper Series, Paderborn University, Faculty of Business Administration and Economics*, (129).

Chen, T. and Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794.

Decroos, T., Bransen, L., Haaren, J. V., and Davis, J. (2019). Actions Speak Louder than Goals. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1851–1861.

Dettmers, T., Pagnoni, A., Holtzman, A., and Zettlemoyer, L. (2024). QLoRA: Efficient Finetuning of Quantized LLMs. In *Proceedings of the Conference on Neural Information Processing Systems*.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of*

the Conference of the North American Chapter of the Association for Computational Linguistics, pages 4171–4186.

Dieter, P., Caron, M., and Schryen, G. (2023). Integrating Driver Behavior into Last-Mile Delivery Routing: Combining Machine Learning and Optimization in a Hybrid Decision Support Framework. *European Journal of Operational Research*, 311(1).

Dietterich, T. G. (2002). Machine Learning for Sequential Data: A Review. In *Structural, Syntactic, and Statistical Pattern Recognition*, pages 15–30. Springer.

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., and Wichmann, F. A. (2020). Shortcut Learning in Deep Neural Networks. *Nature Machine Intelligence*, 2(11):665–673.

Gendreau, M., Hertz, A., Laporte, G., and Stan, M. (1998). A Generalized Insertion Heuristic for the Traveling Salesman Problem with Time Windows. *Operations Research*, 46(3):330–335.

Gupta, A., Dengre, V., Kheruwala, H. A., and Shah, M. (2020). Comprehensive Review of Text-Mining Applications in Finance. *Financial Innovation*, 6(1):1–25.

Hasso, T., Müller, D., Pelster, M., and Warkulat, S. (2022). Who Participated in the GameStop Frenzy? Evidence from Brokerage Accounts. *Finance Research Letters*, 45:102140.

Hausman, W. H. (1969). Sequential Decision Problems: A Model to Exploit Existing Forecasters. *Management Science*, 16(2):B–93.

Hirschberg, J. and Manning, C. D. (2015). Advances in Natural Language Processing. *Science*, 349(6245):261–266.

Hyndman, R. and Athanasopoulos, G. (2018). *Forecasting: Principles and Practice*. OTexts, 2nd edition.

Kapoor, S. and Narayanan, A. (2023). Leakage and the Reproducibility Crisis in Machine-Learning-Based Science. *Patterns*, 4(9).

Kogan, S., Levin, D., Routledge, B. R., Sagi, J. S., and Smith, N. A. (2009). Predicting Risk from Financial Reports with Regression. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 272–280.

Kulinski, S. and Inouye, D. I. (2023). Towards Explaining Distribution Shifts. In *Proceedings of the International Conference on Machine Learning*, pages 17931–17952.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. (2020). BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*.

Merchan, D., Arora, J., Pachon, J., Konduri, K., Winkenbach, M., Parks, S., and Noszek, J. (2022). 2021 Amazon Last-Mile Routing Research Challenge: Data set. *Transportation Science*.

Mishev, K., Gjorgjevikj, A., Vodenska, I., Chitkushev, L. T., and Trajanov, D. (2020). Evaluation of Sentiment Analysis in Finance: from Lexicons to Transformers. *IEEE Access*, 8:131662–131682.

Moraga, P. (2024). *Spatial Statistics for Data Science: Theory and Practice with R*. CRC Press.

Müller, O., Caron, M., Döring, M., Heuwinkel, T., and Baumeister, J. (2021). PIVOT: A Parsimonious End-to-End Learning Framework for Valuing Player Actions in Handball using Tracking Data. In *Proceedings of the Workshop on Machine Learning and Data Mining for Sports Analytics (ECML PKDD)*, pages 116–128.

Ohlmann, J. and Thomas, B. (2007). A Compressed-Annealing Heuristic for the Traveling Salesman Problem with Time Windows. *INFORMS Journal on Computing*, 19:80–90.

Patil, R., Boit, S., Gudivada, V., and Nandigam, J. (2023). A survey of text representation and embedding techniques in nlp. *IEEE Access*, 11:36120–36146.

Pejić Bach, M., Krstić, Ž., Seljan, S., and Turulja, L. (2019). Text Mining for Big Data Analysis in Financial Sector: A Literature Review. *Sustainability*, 11(5):1277.

Raschka, S. and Mirjalili, V. (2017). *Python Machine Learning, 2nd Ed.* Packt Publishing, Birmingham, UK.

Recht, B., Roelofs, R., Schmidt, L., and Shankar, V. (2019). Do ImageNet Classifiers Generalize to ImageNet? In *Proceedings of the International Conference on Machine Learning*, pages 5389–5400.

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. (2019). DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter. *arXiv preprint arXiv:1910.01108*.

Scott, S. L. and Varian, H. R. (2014). Predicting the Present with Bayesian Structural Time Series. *International Journal of Mathematical Modelling and Numerical Optimisation*, 5(1-2):4–23.

Scott, S. L. and Varian, H. R. (2015). Bayesian Variable Selection for Nowcasting Economic Time Series. *Economics of Digitization*, pages 119–136.

Sherstinsky, A. (2020). Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) Network. *Physica D: Nonlinear Phenomena*, 404:132306.

Shumway, R. H. and Stoffer, D. S. (2025). *Time Series Analysis and its Applications*. Springer Texts in Statistics, 5th edition.

Sutskever, I. (2014). Sequence to Sequence Learning with Neural Networks. *arXiv preprint arXiv:1409.3215*.

Tsai, M.-F., Wang, C.-J., and Chien, P.-C. (2016). Discovering Finance Keywords via Continuous-Space Language Models. *ACM Transactions on Management Information Systems*, 7(3):1–17.

Varshney, K. R. (2022). *Trustworthy Machine Learning*. Independently Published.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is All you Need. In *Proceedings of the Conference on Neural Information Processing Systems*, pages 5998–6008.

Verbeek, M. (2004). *A Guide to Modern Econometrics*. John Wiley & Sons, Ltd, 2nd edition.

Wei, L., Zhang, Z., Zhang, D., and Lim, A. (2015). A Variable Neighborhood Search for the Capacitated Vehicle Routing Problem with Two-Dimensional Loading Constraints. *European Journal of Operational Research*, 243(3):798–814.

Wittenbach, J., d'Alessandro, B., and Bruss, C. B. (2020). Machine Learning for Temporal Data in Finance: Challenges and Opportunities. *arXiv preprint arXiv:2009.05636*.

Wooldridge, J. M. (2013). *Introductory Econometrics: A Modern Approach*. South-Western, Cengage Learning, 5th edition.

Xing, F. Z., Cambria, E., and Welsch, R. E. (2018). Natural Language Based Financial Forecasting: A Survey. *Artificial Intelligence Review*, 50(1):49–73.

Zhao, J., Xie, X., Xu, X., and Sun, S. (2017). Multi-View Learning Overview: Recent Progress and New Challenges. *Information Fusion*, 38:43–54.