systems

MDPI

*Article*

# LLM-Driven Sentiment Analysis in MD&A: A Multi-Agent Framework for Corporate Misconduct Prediction

**Yeling Liu †, Yongkang Liu † and Kai Yang ***

College of Economics, Shenzhen University, Shenzhen 518055, China; liuyeling2021@email.szu.edu.cn (Y.L.); 2400128010@mails.szu.edu.cn (Y.L.)

* Correspondence: kaiyang@szu.edu.cn

† These authors contributed equally to this work.

**Abstract**

The textual analysis of Management Discussion and Analysis (MD&A) reveals valuable insights into corporate operational performance and future risks. However, techniques for accurately extracting sentiment from unstructured Chinese MD&A texts still lack comprehensiveness. Existing studies related to sentiment analysis often use lexicon-based methods, which rely on predefined, context-agnostic word lists and accurate Chinese word segmentation and struggle with domain-specific terminology, leading to limited accuracy and interpretability. Although research has attempted to develop context-aware lexicons and language models, these methods still face limitations when applied to long and complex financial texts. To address the limitations, we propose MDARisk, a novel framework for corporate misconduct prediction. The core of MDARisk is a MultiSenti module, which leverages a multi-agent LLM approach to extract comprehensive and contextual sentiment from MD&A. Unlike lexicon methods, our LLM-based module interprets words based on their surrounding semantic context, allowing it to decipher nuanced expressions and specialized financial language. We first conduct an econometric validation using fixed-effects logit models to test whether the MultiSenti-derived MD&A sentiment is significantly associated with subsequent corporate misconduct. We then evaluate out-of-sample predictive utility by adding this sentiment feature to multiple classifiers and assessing its incremental gains over the baseline model. Empirical results demonstrate that our approach provides a more reliable sentiment-based indicator for misconduct risk, achieves higher predictive accuracy, and outperforms the traditional financial sentiment analysis approach. Our MDARisk framework provides a cost-efficient approach for automated disclosure screening, benefiting auditors, regulators, and investors in assessing potential misconduct risks.

**Keywords:** corporate misconduct prediction; sentiment analysis; large language model; MDARisk

check for updates

Academic Editors: He Cai and Maobin Lv

Received: 28 July 2025
Revised: 19 September 2025
Accepted: 21 September 2025
Published: 24 September 2025

**Citation:** Liu, Y.; Liu, Y.; Yang, K. LLM-Driven Sentiment Analysis in MD&A: A Multi-Agent Framework for Corporate Misconduct Prediction. *Systems* **2025**, *13*, 839. https://doi.org/10.3390/systems13100839

**Copyright:** © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

## 1. Introduction

Corporate misconduct refers to deceptive or illegal activities carried out by an individual or a company. Baucus defined corporate illegality as violating administrative and civil law [1], including the falsification of financial statements, environmental violations, and financial misreporting. Studies found that firms' misconduct can significantly undermine market stability, investor wealth, and social trust [2,3]. Energy giant Enron's collapse affected thousands of employees and caused a loss to shareholders as high as USD 74 billion[1]. Therefore, corporate misconduct prediction is highly noticed by regulators and investors to

reduce losses and maintain a healthy market. For external investors and regulators, the financial report is a fundamental way to realize the financial situation and programs, and to discuss the corporate risks and the probability of misconduct. As we know, financial statements are vital components in reports due to their structured and relatively objective data. But it may also present some fraudulent statements that sometimes mislead users. As a section in financial reports, Management Discussion and Analysis (MD&A) conveys textual information such as the potential attitude and opinion from management. Textual information contains sentiment, which reflects the underlying emotional tone or attitude of the text, such as positive, negative, or neutral, and serves as a vital component for forecasting future performance, predicting estate performance, risk and investments [4–7].

While traditional misconduct prediction models often rely on structured financial data, such as accounting ratios and governance metrics [8,9], a growing body of research recognizes that the textual content of corporate disclosures, particularly the MD&A, contains valuable forward-looking information. Accordingly, a growing body of studies has employed sentiment analysis as a crucial tool for extracting managerial tone to predict corporate outcomes, including the likelihood of corporate misconduct [10,11]. However, the effectiveness of this approach hinges on the accuracy of the sentiment measurement itself. Prior dominant sentiment analysis methods in the field of financial text analysis mainly refer to lexicon-based and traditional machine learning (ML) methods, but they face significant limitations in this field. Lexicon methods tally words from predefined lists or dictionaries, such as the Henry List [12] and the Diction, Loughran, and McDonald List [13], to create sentiment indicators to predict misconduct, which are inherently context-agnostic. For instance, the method identifies and counts the word "good", which is in the predefined list, and assigns it a positive weight to construct a positive. These approaches struggle to differentiate sentiment in phrases with identical words and often misinterpret domain-specific terms whose polarity depends entirely on the context. These inaccuracies are particularly detrimental in the context of misconduct prediction, as subtle, nuanced, or deceptively worded phrases may be the important early indicators of underlying issues. In conclusion, the traditional lexicon-based and ML methods perform sentiment analysis from text to misconduct prediction, with several limitations, specifically, context-dependency, domain-specific terms, and cross-lingual barriers. First, they may fail to capture semantic relationships and nuanced expressions [14]. Second, it is hard to update lists frequently and extract the most relevant aspect phrases, regarding the continuous emergence and transferring difficulty of specific words in the specific domain. And a list of characteristic lexicon constructed from the content of foreign annual reports is restricted by an insufficient understanding of Chinese annual reports' textual content [15]. Some words are merely polite, and frequent expressions in the Chinese financial market may be overlooked in the calculation by the traditional method. Table 1 provides specific examples from Chinese MD&A texts (translated into English) that illustrate these challenges. While seemingly general, the inability to correctly interpret such phrases can lead to a misjudgment of sentiment analysis and a firm's true condition, potentially masking the early warning signs of distress that often precede misconduct.

Consider, for instance, a common phrase in MD&A: "The company's performance is better than expected despite challenges" (the first row in Table 1). A traditional lexicon-based approach would process this sentence by simply counting words. It would likely register "better" as a positive term and "challenges" as a negative term. Depending on the specific dictionary weights, the net sentiment could be calculated as neutral or even slightly negative, completely missing the overarching optimistic tone. The method is mechanically blind to the crucial contextual cues provided by phrases like "better than expected" and "despite", which signal resilience and outperformance. Our MDARisk, in contrast, is designed to understand

these relational phrases and syntactic structures. It can recognize that "despite challenges" is a subordinate clause that sets up a contrast, and that the main clause, "performance is better than expected", carries the dominant, positive sentiment. This ability to comprehend complex sentence structures, rather than just isolated words, is precisely what is needed to accurately gauge managerial tone from nuanced disclosures. Table 1 provides further examples.

**Table 1.** Examples of limitations.

| Limitations | Example | Explanation |
| --- | --- | --- |
| context-dependency | The company's performance is better than expected despite challenges. | The overall positive tone may be overlooked. Without the overall context, it is hard to measure the sentiment bias by lists of " challenges" or "better". |
| context-dependency | The term "solid performance" vs. "solid debt" | The neutral adjective "solid" is neutral by itself, yet it may convey opposite sentiments such as positive expressions in expressions like "solid performance" or negative in "solid debt". |
| domain-specific terms | The company earned USD 3 million through hedging business. | Key terms such as "hedging" are domain-specific and not associated with sentiment in general-purpose dictionaries, and the positive verb "earned" may not carry sufficient weight without context. |
| domain-specific terms | We have taken the lead in formulating and developing multiple national standards in the fields of humanoid robots and embodied intelligence, and serve as the deputy head of the "National Humanoid Robot Standards Working Group". | The word "embodied intelligence" is a hot word in the Chinese emerging industry in 2025, and the term "take the lead in" is neutral in some lexicons, although it can be considered that the development trend of the firm may be good in the emerging industry. |

Overall, the lexicon and conventional ML methods have difficulty identifying real emotional tendency with the special language context at a high quality, resulting in the accuracy of misconduct prediction possibly being weakened. Our research aims to enhance the precision of sentiment analysis, subsequently contributing to improved predictive performance in identifying misconduct.

Our study focuses on how to enhance the predictive performance of sentiment. To address limitations, we explore the potential of Large Language Models (LLMs) to enhance sentiment analysis for corporate misconduct prediction. Unlike traditional lexicon methods that rely on predefined word lists [12,13], LLMs are trained on vast textual corpora, enabling them to interpret language in a context-aware manner [16]. This inherent capability allows them to overcome the key challenges identified in Table 1. Specifically, for example, an LLM can differentiate the sentiment of "solid performance" versus "solid debt" by analyzing the entire phrase, and it can infer the sentiment of domain-specific terms like "hedging" from the surrounding sentence structure. We propose that by leveraging these advanced capabilities, a more nuanced and accurate sentiment measure can be extracted from MD&A texts. The central aim of this research is therefore to design and validate a new approach, which we name the MDARisk, which utilizes an LLM-based core, the MultiSenti module, to capture these context-sensitive signals and improve the prediction of corporate misconduct.

To achieve this goal, we conduct a comprehensive empirical study using textual data from the MD&A sections of Chinese A-share listed companies. Our research design follows a two-phase validation strategy. First, we establish the economic significance of our LLM-derived sentiment measure by examining its association with subsequent corporate misconduct. Second, we assess its practical utility by evaluating its incremental contribution to out-of-sample misconduct prediction when compared against both baseline

and traditional lexicon-based models [17]. Our findings reveal that the sentiment extracted by our approach is a powerful and robust predictor of future corporate violations. The inclusion of our MultiSenti-derived feature significantly enhances the performance of predictive models, demonstrating a clear superiority over established financial text analysis methods. These results highlight the substantial value of leveraging advanced language models to identify governance-related risks from corporate disclosures.

This study makes two primary contributions. First, we verify the feasibility and effectiveness of employing a Large Language Model for sentiment analysis in financial texts, which aligns with prior research [16]. Second, our MDARisk offers a lightweight, scalable monitoring tool that complements traditional quantitative methods. From a corporate misconduct prediction perspective, the empirical results provide support for developing and adopting AI-assisted surveillance systems in regulatory technology applications.

The remainder of this article is organized as follows. Section 2 reviews the relevant work on corporate misconduct and sentiment analysis. Section 3 outlines the research methodology of the MDARisk framework. Section 4 details the proposed empirical and experimental settings of the logit regression and out-of-sample test. Our analysis continues with a discussion of our results in Section 5. Finally, Section 6 provides a conclusion to our study.

## 2. Literature Review

### 2.1. Corporate Misconduct and Associated Risks

Corporate misconduct can be broadly defined as deceptive or illegal activities undertaken by firms to secure benefits that outweigh the risks [1,18]. These actions encompass a series of behaviors, including financial misreporting or misrepresentation, corruption, fraud, regulatory violations, and other misconduct [18–20]. Corporate misconduct represents a series of governance risks and even impacts on the financial market and society. Researchers have shown that it triggers significant negative market reactions, with cumulative abnormal returns averaging around −4.1% following media disclosure [2]. These risks arise from uncertain activities such as the company's operating environment, organizational activities, management practices, and subjective decision-making, which may potentially lead to financial losses. It is commonly believed that corporate risk can be assessed from both internal and external perspectives, reflecting the impact of corporate governance and the external environment.

Research from an internal governance perspective has shown that the main influencing factors are the firm's size and financial status [21,22]. The larger firm has difficulty maintaining the monitoring effectiveness of its sub-units, which increases the probability of misconduct occurring. Also, firms may find alternative sources through illegal means to mitigate low profitability, which results in an abnormal capital structure (e.g., a high debt ratio). Hence, we control for firm size and debt ratio in econometric validation. Typically, companies' characteristics, such as the number of board members, have also been found to influence corporate misconduct. Smaller and more diverse boards are generally associated with stronger internal controls and monitoring, which can reduce the likelihood of misconduct [23]. Factors such as the dual role of a chairman who also serves as the CEO and high executive compensation incentives can significantly reduce the risk of corporate misconduct [24]. Further research has found that management's use of related-party transactions for personal gain can exacerbate risks of corporate misconduct, and overconfidence in managers increases the level of risk the company faces [25]. Thus, we control for equity composition in econometric validation.

From an external perspective, research has shown that short-selling mechanisms can help reduce corporate misconduct tendencies and increase the likelihood of violations being detected [26]. Other studies indicate that capital market openness, reflected in the proportion of equity held by international investors who are more concentrated on firm

governance, can effectively curb corporate misconduct [27]. The annual report still remains the most vital source to analyze the condition and risks of a company, and recent literature increasingly recognizes that the unstructured text of the Management Discussion and Analysis (MD&A) contains crucial, non-quantitative precursors to misconduct [10,28,29]. The MD&A serves as a channel for managers to communicate their perspective, and the decisions they make—whether consciously or subconsciously—can indicate underlying issues well before these concerns become apparent in the financial statements. These early indicators may provide crucial clues for investigating the potential for misconduct in subsequent years.

Current research identifies the misconduct risk from the MD&A by analyzing various linguistic signals. For instance, some studies focus on textual readability, finding that managers intending to conceal poor performance may use overly complex language, which is correlated with a higher likelihood of financial restatements [4,6]. Other approaches analyze specific language patterns thought to be indicative of deception [30]. Among these signals, managerial sentiment or tone has emerged as a particularly potent indicator, as it can reflect deteriorating business conditions, internal control weaknesses, or managerial pressure—all known risk factors for misconduct [28,31].

Therefore, while the MD&A does not contain explicit confessions of misconduct, it provides a rich tapestry of linguistic signals, especially the managerial sentiment, which is currently used as a valuable proxy to assess a firm's underlying risk of misconduct. Our study builds directly on this foundation, contributing by focusing on and significantly improving the accuracy of managerial sentiment.

### 2.2. Sentiment Analysis in Finance

Sentiment analysis has become a cornerstone of modern financial research, providing a systematic way to quantify the valuable, non-numerical information embedded in textual disclosures [32]. The sentiment features extracted from text, like sentiment bias, tone and readability, provide a new perspective on classic research of understanding a firm's quality of financial data, future performance, risk prediction, and investments [4–6]. For instance, studies indicate that tone changes in MD&A can predict investor behavior [33]. The primary methods for conducting sentiment analysis in finance have evolved significantly over time, progressing from simple word-counting techniques to sophisticated neural network models.

The foundational approach is the lexicon-based method. This method needs to split the text into words and count the frequency of words by associating them with specific word lists. Commonly used word lists in the research include Henry [12], Harvard General Inquirer, Diction, and Loughran and McDonald [13], which categorize features such as positive, negative, uncertain, readability, etc. Previous research has shown that word lists (e.g., Henry, Diction, and Loughran and McDonald), which are set for finance and accounting, have a good measurement effect in sentiment analysis [30,34]. The primary advantage of this method is its simplicity, transparency, and reproducibility. But some studies have shown that the Henry lexicon, when used for sentiment analysis in MD&A sections of annual reports, does not show significant relationships between textual sentiment and company performance [35]. This may be because predefined lexicons lack sentiment words relevant to different industries, and words labeled as negative may not necessarily convey negative connotations in financial documents.

A more advanced approach to analyze sentiment is traditional machine learning (ML), which employs algorithms like Linear Regression, Support Vector Machines, and Random Forests to parse data, learn from it, and make informed decisions or predictions. It requires significant amounts of structured or labeled data for training algorithms effectively and

then applies algorithms to the entire text corpus [36]. Unlike lexicon methods, ML models can learn the importance of word combinations and other features from the data itself. However, the predictive performance of traditional ML methods is highly dependent on the availability of large, high-quality, manually labeled datasets for training, which can be a significant bottleneck [37,38].

The rise of deep learning (DL), related to the rapid advancement of artificial intelligence hardware, marked a further evolution [39]. As a subset of ML, DL models, particularly those using multi-layered Artificial Neural Networks (ANNs), can automatically learn complex, hierarchical features directly from raw text without extensive manual feature engineering. These models have shown strong performance in various NLP tasks, including sentiment analysis [28]. Nevertheless, they require very large datasets (often millions of data points) and high-performance computing to train, leading to limited applications to misconduct prediction [28].

More recently, the emergence of Large Language Models (LLMs), such as the GPT series, has represented a breakthrough and a paradigm shift in natural language processing [16,40]. Pre-trained on vast, internet-scale text corpora, LLMs possess a remarkable ability to understand grammar, context, and nuance. Unlike traditional models that required task-specific training, LLMs are generative and solve the task of predicting the next token based on preceding tokens, much like an autocomplete [16]. Therefore, LLMs can perform sentiment analysis in a zero-shot or few-shot setting, simply by following instructions in a prompt. Researchers have begun to employ ChatGPT-3.5 and ChatGPT-4 to examine whether tone manipulation occurs in the MD&A. Song et al. [40] suggest that tools like ChatGPT can help identify and reduce tone manipulation in financial texts.

This methodological evolution from fixed dictionaries to context-aware generative models highlights the potential of sentiment analysis in financial text research, providing a foundation for predicting corporate misconduct.

### *2.3. Sentiment Analysis and Corporate Misconduct*

Recent research has applied various sentiment analysis methods to the prediction of corporate misconduct. Campbell and Shang's research [10] demonstrated that information extracted from employee reviews can be used to develop measures with useful properties for measuring misconduct risk. Sentiment provides a direct and clear reflection of companies' operational and financial status [41]. Some studies have shown that sentiment in the MD&A has the predictive ability of the financial distress model [15]. The increased negative sentiment in the text reflects concerns about the company's performance. Following research by Guo et al. [31], negative MD&A sentiment enhances financing costs and problems due to the risk aversion of financial institutions such as banks, leading to exacerbation of the company's financial distress, and further raises the risk of company misconduct.

However, while the theoretical link is strong, its practical utility in prediction models is entirely dependent on the accuracy of the initial sentiment measurement. When analyzing the MD&A—a document often crafted to obscure as much as it reveals—the limitations of traditional sentiment analysis methods become particularly acute.

The first major challenge is context-dependency. Lexicon-based methods, being context-agnostic, are easily misled by nuanced financial language. Consider the phrase (seen in the first row of Table 1), "The company's performance is better than expected despite challenges." A lexicon-based tool would mechanically register "better" (positive) and "challenges" (negative), potentially calculating a neutral score and missing the overarching optimistic tone. This inability to understand relational phrases and syntactic structures is a critical failure, especially in the misconduct context, where managers may use such complex sentences to deliberately manage impressions and mask underlying issues [42,43].

The second challenge is the prevalence of domain-specific terms. Compared to general text, MD&A text includes various industries' potential information that is contained in a large amount of domain-specific terms, which is hard to handle with fixed lexicons. Words like "liability" and "hedging" are sentiment-neutral in both general dictionaries and finance-specific dictionaries but carry significant, context-dependent weight in a financial report. Even context-aware lexicons, such as the one proposed by Kumar and Uma [44], which are improved by incorporating semantic similarity and dynamic weight adjustments, are still limited in handling highly technical and emerging financial terms. For instance, words like "stablecoin" or "depeg" may not appear in context-aware lexicons at all, although they convey strong risk signals when mentioned in financial disclosures. The rise of machine learning methods, such as the use of BERT models, has significantly advanced context-aware sentiment analysis in short finance text [45,46]. Despite these advancements, they still face the challenge when processing long and complex texts with input limitations, such as a maximum of 512 tokens in typical BERT, which makes it challenging to process longer contexts in MD&A without losing coherence [47]. This leads to an underestimation of risk signals embedded in technical discussions.

These issues are further compounded by the unique characteristics of the Chinese language and disclosure environment. Lexicon methods, many of which are direct translations from English, struggle with the inherent ambiguity of Chinese word segmentation [37,48]. Specifically, the translation-based lexicons often fail to understand Chinese-specific expressions and idioms, leading to mismatches between the lexicon and actual text. In addition, researchers have also noted that context-dependency and semantic ambiguity in Chinese, even the basic task of separating words, are difficult, leading to lower accuracy for traditional word list methods [48]. An effective analysis tool must be able to cope with these questions to capture the true managerial sentiment.

In conclusion, accurately predicting misconduct from MD&A text requires a tool that can navigate complex sentence structures and interpret domain-specific vocabulary in context. The inherent limitations of conventional lexicon and machine learning methodologies render them inadequate for this task, resulting in a noisy and unreliable sentiment signal. This analytical gap provides a strong motivation for exploring more advanced, context-aware technologies like LLMs, which form the basis of our proposed approach. When leveraging LLMs, they provide new possibilities for addressing the deep challenges in MD&A sentiment analysis through their powerful contextual understanding capabilities. In particular, LLMs can perform sentiment analysis of text without needing additional training and solve tasks based on their basic reasoning and world understanding— capabilities that are difficult to achieve with conventional NLP techniques, which face the limitations mentioned above [16].

## 3. Research Design and Methodology

Our methodology follows the DSR paradigm, structured in three core phases: problem identification and design (Section 3.1), artifact implementation (Section 3.2), and rigorous evaluation (Section 3.3). This approach guides the development of our end-to-end framework, MDARisk, for misconduct prediction.

### *3.1. Problem Identification and Design Requirements*

Our research methodology is inspired and guided by Hevner et al.'s principles of Design Science Research (DSR), which focus on creating and evaluating innovative artifacts to solve practical organizational problems [49]. We adopt this approach to systematically design and validate our primary artifact: the MDARisk, an integrated system for corporate misconduct prediction.

#### 3.1.1. Problem Identification and Motivation

As established in the Introduction and detailed in Section 2.3, the problem we address is well-documented: traditional lexicon-based sentiment analysis methods are often ineffective and inaccurate for Chinese MD&A texts. This deficiency stems from their inability to process long sentences and to capture the real meanings of highly context-dependent words and domain-specific terms [13,15,37]. This limitation significantly hinders the use of textual sentiment for predicting corporate misconduct.

#### 3.1.2. Objectives of the Artifact

Regarding the limitation of inaccurately measuring the sentiment of MD&A text for predicting downstream misconduct, our objective is to develop MDARisk, whose core innovation is a multi-agent sentiment component, the MultiSenti module (step 2), designed to extract contextual sentiment with higher accuracy, thereby improving the overall predictive performance for corporate misconduct. MDARisk is designed to overcome the limitations of existing methods by leveraging the advanced natural language understanding capabilities of Large Language Models.

#### 3.1.3. Design Requirements for the MDARisk Artifact

Based on the challenges identified in our literature review and the practical requirements of corporate misconduct prediction, we translated our high-level objectives into three concrete design requirements (DRs) for the MDARisk artifact. These requirements guided the technical design of our MultiSenti module and established clear criteria for its subsequent validation.

(i) DR1: Capability to Interpret Context-Dependent and Domain-Specific Language.

Prior work shows that traditional lexicons misclassify polarity in finance and struggle with context and domain terminology [13,15,37]. An effective artifact must therefore be able to derive sentiment from the holistic semantic meaning of a sentence, not from isolated keywords. Therefore, our primary requirement is for the artifact to perform sentiment analysis in a context-aware manner, accurately interpreting words based on their surrounding text so that it can analyze the sentiment in MD&A much more accurately and comprehensively.

In order to meet this requirement, we implement a Large Language Model (ChatGLM4-Flash) as the core analytical engine. Unlike lexicon lookups, an LLM's architecture inherently processes language sequentially, enabling it to understand syntactic relationships and contextual nuances based on its vast pre-training on diverse corpora [16]. To further hone this capability for our specific task, we employ a structured prompt that directs the model to focus on financial signals.

The success of this design choice is empirically evaluated in our Phase 2 out-of-sample tests (Section 5.4), where the superior predictive performance of our approach over traditional lexicon methods directly demonstrates its enhanced contextual understanding.

(ii) DR2: Generation of Stable and Reproducible Outputs.

LLMs exhibit stochastic behavior in their generation process, which poses a challenge for reproducible empirical research. Ensuring output stability is therefore a critical design specification. To meet this requirement, we draw upon principles of ensemble methods, which are known to improve model robustness and consistency in machine learning and NLP [50].

Our MultiSenti module implements this by employing a multi-agent procedure: for each MD&A text, three independent analytical agents are run. The final sentiment label is then determined by a majority vote, and the continuous score is derived from the arithmetic mean of the individual outputs. This ensemble approach effectively mitigates random variation and produces a stable, reliable sentiment metric suitable for rigorous research.

The fulfillment of this specification is architecturally embedded within the MultiSenti module (see Algorithm 1), and its robustness is further confirmed through repeated runs in our machine learning evaluation.

(iii) DR3: Compatibility with Standard Empirical Research Pipelines.

For any novel textual feature to be practically useful in finance and accounting research, it must be easily integrable into established analytical workflows [50]. This necessitates a design that produces outputs in a standard, machine-readable format. Accordingly, we specified that the artifact must generate outputs that are directly compatible with common statistical and machine learning packages.

The MultiSenti module is engineered to produce two such outputs for each document: a binary sentiment label (0 or 1) and a continuous, normalized sentiment score (from 0.0 to 1.0). These simple numerical formats allow the sentiment metric to be used as an independent variable in econometric models and as a feature in ML classifiers without requiring complex transformations.

We provide direct evidence of this specification's fulfillment by seamlessly incorporating these outputs into our fixed-effects logit models (Section 5.2) and the suite of five distinct ML classifiers (Section 5.4).

Figure 1 presents the MDARisk. Step 2 is instantiated by the MultiSenti module. Algorithm 2 operationalizes MDARisk as a whole: it calls MultiSenti to obtain {label, avg_sentiment_score} for each MD&A, assembles features with structured controls, performs a time-based split with SMOTE on the training set, and trains/evaluates the classifiers and fixed-effects models. Thus, Algorithm 1 is the pseudocode for the end-to-end artifact (MDARisk), while MultiSenti is the core subroutine implementing our multi-agent sentiment design.

**Algorithm 1**: The MultiSenti Module.

Input:
*text*: A single MD&A text document (string).
*N_runs*: Number of independent analysis runs (integer, default = 3).
*LLM_model*: A pre-defined Large Language Model (e.g., ChatGLM4-Flash).
*Prompt*: A structured prompt template defining the role, task, and output schema.
Output:
*final_sentiment*: The aggregated sentiment label (binary: 0 or 1).
*avg_score:* The averaged sentiment intensity score (float: 0.0–1.0).
Procedure:
*sentiment_outputs* ← []
*score_outputs* ← []
*FOR i = 1 TO N_runs DO:*
# Each run is an independent agent
*raw_output* ← *LLM_model(Prompt + text)*
*parsed_output* ← *ParseJSON(raw_output)*
*sentiment_outputs.append(parsed_output.sentiment)*
*score_outputs.append(parsed_output.score)*
*END FOR*
# Aggregate results from all agents
*final_sentiment* ← *MajorityVote(sentiment_outputs)*
*avg_score* ← *Average(score_outputs)*
*RETURN (final_sentiment, avg_score)*

**Algorithm 2**: The MDARisk End-to-End Pipeline for Misconduct Prediction.

Input:
*D_text*: A collection of raw MD&A texts for each firm-year.
*D_financials*: A dataset of raw financial variables and misconduct labels for each firm-year.
Classifiers: A list of machine learning models (LogisticRegression, SVM, GBDT, XGBoost).
Output:
*ML_Results*: A dictionary containing performance metrics for each trained classifier.
*FE_Results:* A dictionary containing fixed-effects logit regression outputs.
Procedure:
1: # --- Step 1: Data Acquisition and Preprocessing ---
2: *Cleaned_text* ← *Clean(D_text)*
3: *Processed_financials* ← *Process(D_financials)*
4: *Data* ← *Merge(Cleaned_text, Processed_financials)*
5: # --- Step 2: Multi-Agent LLM Sentiment Quantification ---
6: *For EACH report IN Data DO:*
7: # Call the core sentiment module to get aggregated label and score
8: *(report.sentiment, report.avg_sentiment_score)* ← *MultiSenti(report.text, N_runs = 3)*
9: *END FOR*
10: # --- Step 3: Feature Set Construction ---
11: *Control variables* ← *Select(Data, columns = ['Size', 'LEV', 'Board', 'Top1', 'InsInvestorProp'])*
12: *Sentiment_Feature* ← *Select(Data, columns = ['avg_sentiment_score'])*
13: *X* ← *Combine(Control_Variables, Sentiment_Feature)*
14: *Y* ← *Data.Misconduct*
15: # --- Step 4: Model Training, Evaluation, and Comparison ---
16: # Phase 1: Econometric Validation
17: *FE_Model* ← *FixedEffectsLogit(formula = "Y_Sentiment_Feature + Control_Variables + FirmFE + YearFE")*
18: *FE_Results* ← *FE_Model.fit(data=Data)*
19: # Phase 2: Predictive Validation
20: *ML_Results* ← *Train_and_Evaluate_Classifiers(Classifiers, X, Y)*
21: *RETURN (FE_Results, ML_Results)*

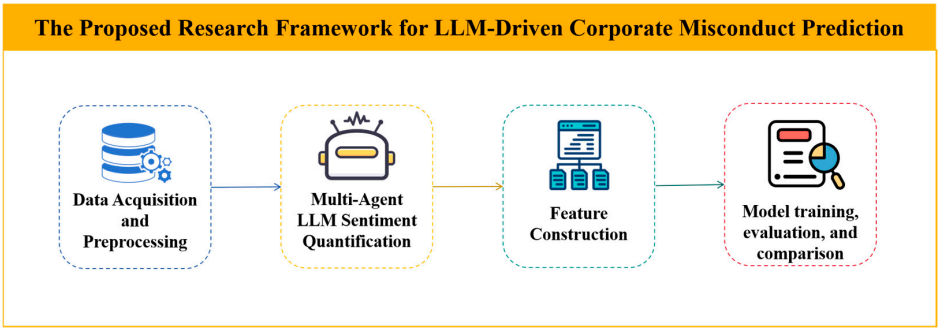



**Figure 1.** The Architecture of the MDARisk for Corporate Misconduct Prediction.

### *3.2. Artifact Implementation and Feature Engineering*

Guided by the design requirements established above, we proceeded to design and implement the MDARisk framework. This section details the core components and technical implementation of our artifact, corresponding to the "build" phase of the design science cycle.

#### 3.2.1. Data Sources and Preprocessing

The foundation of this study rests upon two parallel streams of high-quality data: unstructured MD&A texts from corporate annual reports and structured firm-level data encompassing financial indicators, governance variables, and regulatory misconduct records. As illustrated in Figure 2, both data streams undergo rigorous preprocessing to ensure integrity and accuracy. For textual data, this includes cleaning procedures to remove non-linguistic artifacts. For structured data, standard practices such as missing value imputation and outlier handling are applied. The detailed description of our sample, specific data sources, and the full preprocessing workflow are provided in Section 4.1.

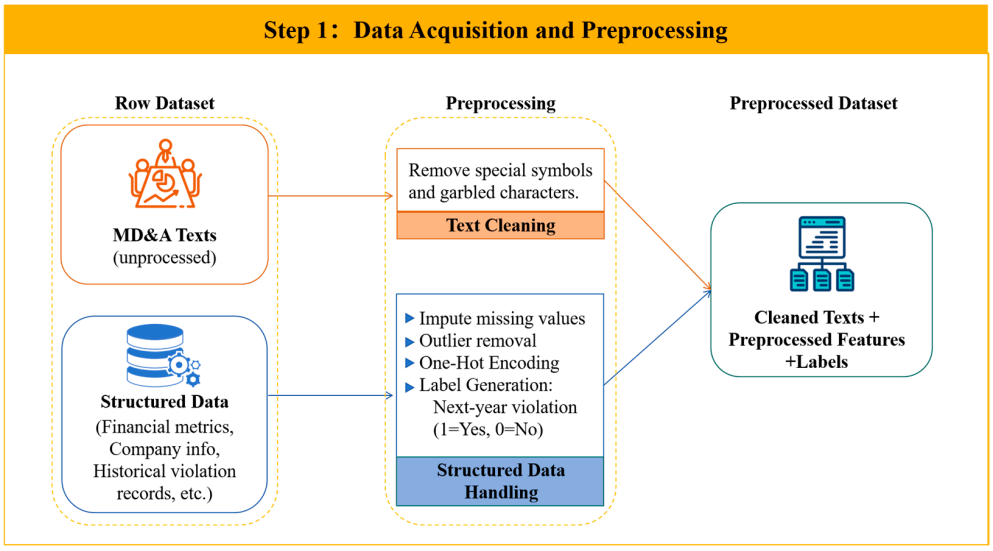



**Figure 2.** Step 1: Data acquisition and preprocessing.

#### 3.2.2. The MultiSenti Module: A Multi-Agent Approach to Sentiment Quantification

The core technical innovation of our MDARisk is the MultiSenti module, which is designed to accurately quantify sentiment from complex financial texts. This module replaces traditional dictionary-based methods with a sophisticated multi-agent LLM approach, as depicted in Figure 3.

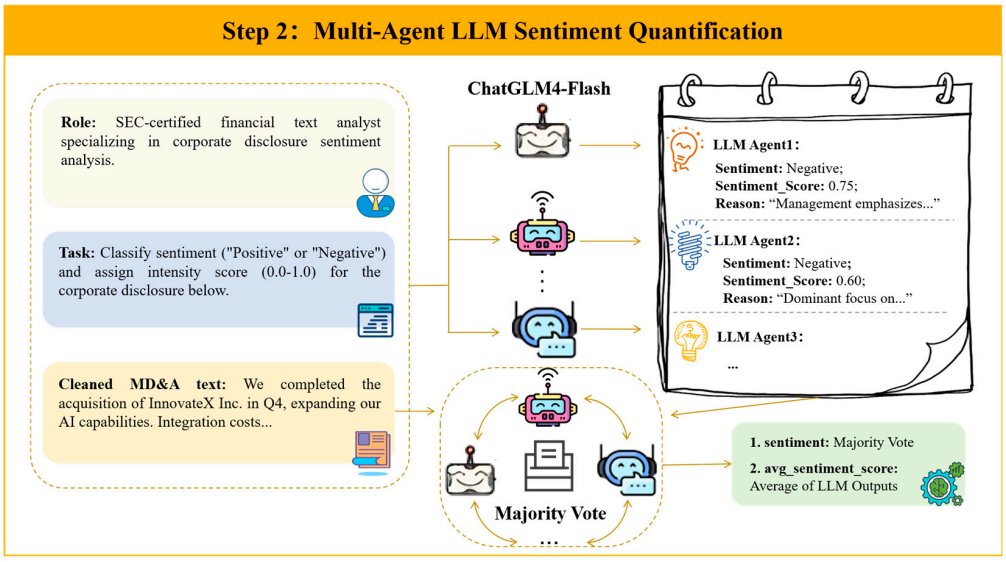



**Figure 3.** Step 2: The Workflow of the MultiSenti Module.

Traditional dictionary-based methods often fail to capture the context, tone, and nuances inherent in financial discourse, as discussed above. To more accurately quantify sentiment, this study innovatively employs Zhipu AI's ChatGLM4-Flash, a Large Language Model noted for its proficiency in understanding complex Chinese texts.

We selected ChatGLM4-Flash for several key reasons. First, as a model developed by a leading Chinese AI company (Zhipu AI), it has been specifically optimized for understanding complex Chinese texts, including the formal and often nuanced language found in financial reports. This makes it particularly well-suited for our research context compared to models primarily trained on English data. And its "Flash" version offers an excellent balance between high performance and computational efficiency, enabling us to process a large corpus of annual reports within a feasible timeframe and budget, a crucial consideration for the scalability of our framework. Finally, preliminary testing indicated its strong capabilities in zero-shot classification tasks, which is ideal for our application, where fine-tuning is not performed.

A critical element of the MultiSenti module is our structured prompt design. Instead of simply asking for sentiment, we engineer the prompt to elicit expert-level analysis, which contains three parts:

- Role: "SEC-certified financial text analyst specializing in corporate disclosure sentiment analysis."
- Task: "Classify the provided MD&A excerpt as Positive or Negative and assign an intensity score in [0.0, 1.0], where 0.0 is most negative and 1.0 is most positive."
- Output schema and rationale cue: "Return JSON with fields {'label', 'score', 'rationale'}, where 'rationale' briefly cites key phrases driving the decision (e.g., risk disclosures vs. forward-looking achievements)."

This schema encourages contextualized reading and ensures parsable outputs. The rationale text is not used as a feature but helps audit the classification and reduces off-task generation.

To ensure stable outputs (DR2), MultiSenti runs three independent analytical instances ("agents") of the same model on the same text. Three agents provide the minimum robust majority vote while keeping computational cost modest. Each agent returns a label and a score; the final label is determined by majority vote, and the final intensity score (avg_sentiment_score) is the arithmetic mean of the three scores. This ensemble style aggregation materially improves robustness and consistency for empirical analysis, making it suitable for rigorous empirical analysis.

#### 3.2.3. Feature Construction

To construct a comprehensive feature set, we integrate our novel textual sentiment metric with a set of well-established firm-level control variables. This feature construction pipeline is illustrated in Figure 4. The primary feature of interest, *avg_sentiment_score*, is derived from MultiSenti as detailed in Section 3.2.2.

To isolate the effect of this sentiment metric and mitigate omitted variable bias, we incorporate a series of control variables selected based on the established literature in corporate finance and accounting [10,21,28]. These controls capture key dimensions of a firm's financial health, governance, and ownership structure. This methodology enables a direct assessment of whether sentiment provides significant predictive information beyond traditional indicators. The precise definitions of all variables used in our models are presented in Section 4.2 and Table 2.

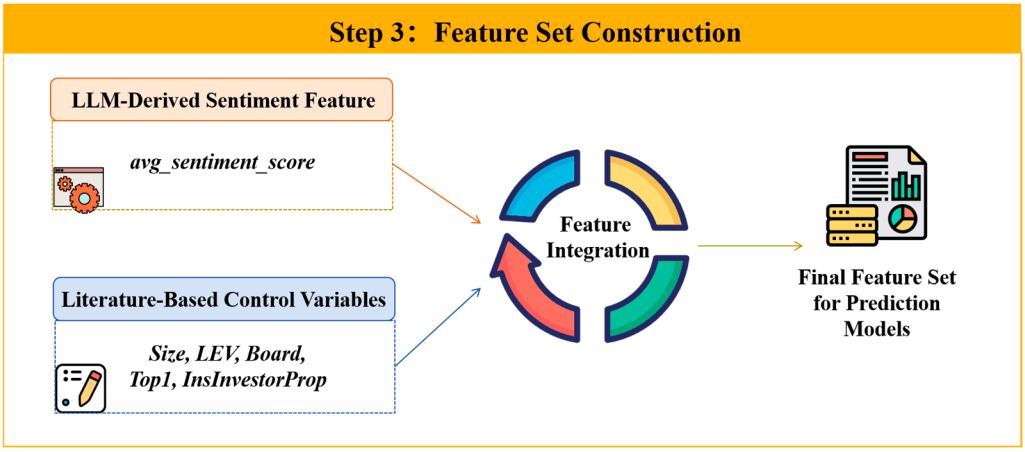



**Figure 4.** Step 3: Feature set construction.

**Table 2.** Variables and measurements.

| Variables | Measurements | Reference & Sign |
|---|---|---|
| Misconduct | When a listed company is found to have violated regulations during an audit, Misconduct = 1; otherwise, Misconduct = 0 | |
| avg_sentiment_score | Refer to 4.1 | – |
| Size | The natural logarithm of the total assets | [20,48,51]; – |
| Board | The natural logarithm of the total number of board members | [21,43,52]; – |
| LEV | Total Liabilities/Total Assets | [20,48,51]; + |
| Top1 | The proportion of equity held by the company's largest shareholder | [21,43,52]; + |
| InsInvestorProp | The proportion of equity held by institutional investors in the company | [21,43,52]; – |

Notes: The sign in column 3 represents the possible expected impact of independent variables on the dependent variable.

### *3.3. Evaluation Strategy: Verification and Validation*

As the final stage of the design science cycle, we follow Hevner et al.'s principle of rigorous evaluation [49] and develop a comprehensive evaluation plan, incorporating both technical verification and a two-phase empirical validation to assess the artifact's effectiveness from different perspectives.

#### 3.3.1. Technical Verification of Design Requirements

First, we perform a technical verification to ensure the artifact adheres to its specified design. DR2 (Output Stability and Reliability) is addressed fundamentally by the design of the MultiSenti module itself. The use of a multi-agent procedure with majority voting and score averaging is an explicit mechanism to mitigate the stochasticity inherent in LLMs and ensure consistent outputs [53]. The overall robustness of our findings is further reinforced by averaging the performance metrics across multiple runs with different random seeds in our machine learning evaluation (Phase 2 Validation). Simultaneously, DR3 (Seamless Integration) is directly verified by the successful implementation of our entire research pipeline. The dual outputs from MultiSenti—the binary sentiment label and the normalized continuous score—were used directly as features in both the fixed-effects logit models for our Phase 1 Validation (Sections 5.2 and 5.3) and the suite of five distinct machine

learning classifiers in our Phase 2 Validation (Section 5.4) without requiring any complex pre-processing or transformation. This demonstrates the practical compatibility of our artifact with standard econometric and machine learning workflows.

#### 3.3.2. Two-Phase Validation Strategy

Next, we execute our two-phase validation strategy to systematically assess our artifact, with each phase directly corresponding to the validation of our design requirements, as shown in Figure 5.

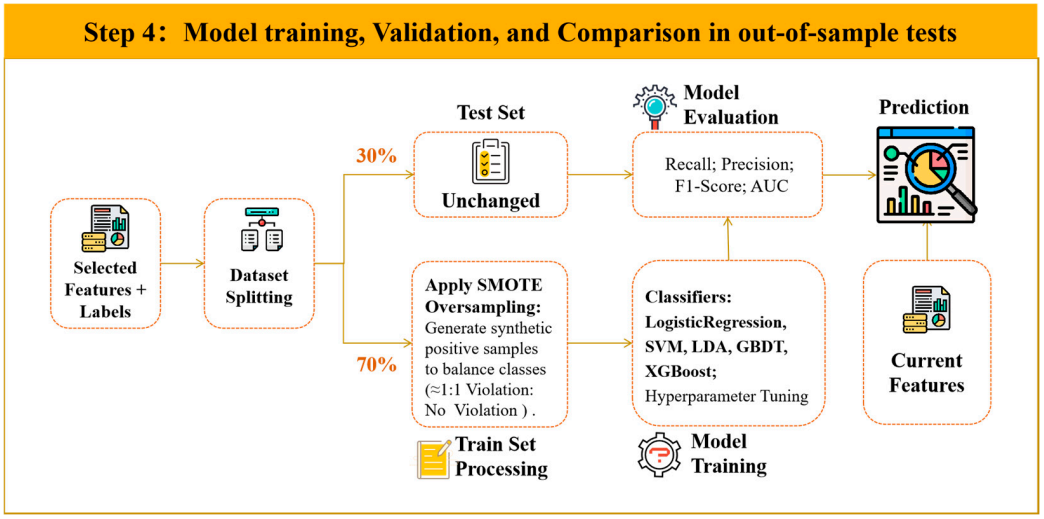



**Figure 5.** Step 4: Model training, validation, and comparison in out-of-sample tests.

The first phase is an econometric validation designed to primarily validate DR1 from an economic significance perspective. The primary goal of this phase is to establish the construct validity of our core MultiSenti module by proving that the sentiment it extracts is not just statistical noise but a meaningful economic signal related to governance risk. Therefore, we employ a fixed-effects logit model to formally test our hypothesis (H1) that the MultiSenti-derived sentiment score is significantly associated with subsequent corporate misconduct. A significant result would provide strong evidence that our artifact successfully captures nuanced, context-aware information that is relevant to a firm's underlying risk profile, thereby fulfilling a core aspect of DR1. The detailed protocol for this phase is presented in Section 4.5.

The second phase is a predictive validation that evaluates the practical utility of the entire MDARisk. This phase assesses the incremental value of the complete MDARisk in the misconduct prediction task. The objective is to solve the binary classification problem of predicting whether a firm will commit a regulatory violation in the following year. To achieve this, we build, train, and compare multiple machine learning models using a time-based holdout split and SMOTE to address class imbalance [54]. To rigorously assess the incremental predictive value, we compare three feature configurations: a baseline model, a model augmented with traditional sentiment, and a model augmented with our MultiSenti-derived sentiment score.

Our evaluation in this phase focuses on metrics suited for imbalanced datasets, such as Recall, F1-Score, and the Area Under the ROC Curve (AUC), to provide a comprehensive assessment of model performance [55,56]. The complete experimental design is detailed in Section 4.6.

## 4. Datasets and Validation Protocol for MDARisk

This section details the data, variables, and experimental protocols used to implement and evaluate the MDARisk framework introduced in Section 3. We first describe the sample and data sources, then outline the procedures for our two-phase validation: (i) an econometric test to validate the construct validity of the MultiSenti module and (ii) an out-of-sample predictive evaluation to validate the practical utility of the complete MDARisk.

### *4.1. Sample and Data Sources*

We study A-share non-financial firms from 2019 to 2023. Annual report MD&A texts are obtained from Cninfo; firm characteristics and governance variables come from CSMAR. We exclude delisted firms and observations with missing key fields, yielding 19,988 firm-years.

### *4.2. Implementing MultiSenti to Measure Sentiment Score of MD&A*

We implemented the MultiSenti module as described in Section 3.2.2, using Zhipu AI's ChatGLM4-Flash to analyze the MD&A sections of corporate annual reports. For each MD&A text, the module performs three independent analytical runs to generate a final majority-vote sentiment label and an averaged intensity score (avg_sentiment_score). Considering the lagged effect of disclosure sentiment on future misconduct, the MD&A texts are from the 2018–2022 period, corresponding to misconduct outcomes from 2019 to 2023.

### *4.3. Measure of Corporate Misconduct*

We measure corporate misconduct using regulatory violation data from the CSMAR database. To be more specific, we exclude violations that resulted from personal actions of the company's shareholders or management. Our primary dependent variable, *Misconduct*, is a dummy variable that equals 1 if a firm was sanctioned for a regulatory violation in the year following the annual report's disclosure and 0 otherwise. The period for this variable ranges from 2019 to 2023, reflecting the early warning effect of sentiment in the annual report texts.

### *4.4. Feature Construction and Controls*

To mitigate omitted variable bias and isolate the incremental effect of managerial sentiment, we include a set of control variables widely recognized in the corporate finance and accounting literature as significant predictors of corporate misconduct. Following prior studies [20,48,51], we control for firm size (Size) and financial leverage (LEV), as larger and more leveraged firms may face different levels of scrutiny and financial pressure. We also account for corporate governance characteristics [21,43,52], including board size (Board), ownership concentration (Top1), and the proportion of institutional investors (InsInvestorProp), as these factors are known to influence managerial oversight and firm behavior [21,57]. Variable definitions follow Table 2; this step corresponds to MDARisk Step 3 in Figure 1.

### *4.5. Econometric Validation Protocol (Phase 1)*

As the first phase of our validation strategy, this econometric analysis aims to establish the construct validity of our MultiSenti module by formally testing its output against a key design requirement. Specifically, this phase validates DR1 (Capability to Interpret Context-Dependent and Domain-Specific Language) by examining whether the context-aware sentiment score generated by our artifact is economically meaningful.

To do so, we formulate our primary hypothesis, which posits that a more negative sentiment in the MD&A—as captured by a lower avg_sentiment_score—is associated with a higher probability of subsequent corporate misconduct. The hypothesis is as follows:

**H1:** *There is a significant positive correlation between the negative sentiment tendency in MD&A text and the risk of corporate misconduct.*

Through testing this hypothesis, we can verify if our artifact captures a genuine signal of governance risk rather than statistical noise. A significant finding would provide strong evidence that MultiSenti, by fulfilling DR1, can extract information relevant to corporate risk that is often missed by context-agnostic methods. This step is a prerequisite for the subsequent predictive validation in Phase 2.

Based on H1, we construct a panel logit model as a baseline model to examine the relationship between MD&A sentiment and the likelihood of corporate misconduct. The use of a logit model is standard for rare-event corporate outcomes [58]. To control for unobserved firm-specific heterogeneity and common time shocks, we incorporate both firm and year fixed effects in our preferred specification.

$$\begin{aligned} Pr(Misconduct = 1) = \alpha_0 + \beta_1\, avg_sentiment_score + \\ \beta_2\, Size + \beta_3\, Board + \beta_4\, LEV + \beta_5\, Top1 + \beta_6\, InsInvestorProp + \lambda_{i.firm} + u_{i.year} + \varepsilon_{it} \end{aligned} \quad (1)$$

The dependent variable is the corporate misconduct risk indicator, which is measured by a dummy variable representing whether the company has been penalized for violations. The primary independent variable is the sentiment score of the annual report text. Control variables include company size, board size, debt leverage, ownership concentration, and institutional concentration. To prevent the influence of corporate misconduct on the independent variables, all independent variables are lagged by one period. The approach concentrates on the impact of ex ante factors on the company's motivations and severity of violations.

### *4.6. Predictive Validation Protocol (Phase 2)*

To operationalize the second phase of our validation, which evaluates the predictive utility of the full MDARisk, we design a series of out-of-sample prediction experiments using multiple machine learning classifiers: Logistic Regression, Support Vector Machine (SVM), Gradient Boosting Decision Trees (GBDTs), Linear Discriminant Analysis (LDA), and XGBoost. Using multiple classifiers ensures that our findings are robust and not contingent on the specific assumptions of a single algorithm, a common practice in predictive modeling research [59].

We compare three feature configurations:

- Baseline Features: Includes firm-level controls such as industry classification, institutional ownership ratio, etc.;
- Baseline + Traditional Sentiment: The baseline features augmented with sentiment scores generated by the lexicon-based method of Jiang et al. [17];
- Baseline + LLM Sentiment: The baseline features augmented with the *avg_sentiment_score* produced by MultiSenti.

The MultiSenti module generates the sentiment feature by averaging the scores from three independent analytical runs for each MD&A document. To ensure a realistic forecasting scenario, we use a time-based holdout split (70% for training, 30% for testing) and apply SMOTE to the training set only to handle the class imbalance of the minority misconduct class [52]. Model performance is assessed using metrics appropriate for imbalanced datasets, including Accuracy, Recall, F1-score, and AUC. All results are averaged across multiple runs with different random seeds to ensure robustness.

## 5. Validation Results

This section presents the results of the two-phase validation strategy designed to evaluate our MDARisk framework. We begin with descriptive statistics, followed by the

results of our Phase 1 econometric validation, which assesses the construct validity of the core MultiSenti module. We then present the results of our Phase 2 predictive validation, which demonstrates the practical utility of the complete MDARisk framework.

### *5.1. Descriptive Statistics*

Table 3 shows the descriptive statistics. The mean of Misconduct is 0.1251, indicating that approximately 12.5% of the firm-year observations in our sample are associated with a corporate misconduct event in the subsequent year.

**Table 3.** Descriptive Statistics.

| VarName | Obs | Mean | SD | Min | Max |
|---|---|---|---|---|---|
| Misconduct | 19,988 | 0.1251 | 0.331 | 0.0000 | 1.0000 |
| Sentiment | 19,988 | 0.9510 | 0.216 | 0.0000 | 1.0000 |
| avg_sentiment_score | 19,988 | 0.7967 | 0.129 | 0.2300 | 0.9500 |
| Size | 19,988 | 22.3329 | 1.358 | 17.5453 | 28.6969 |
| Board | 19,988 | 2.0908 | 0.197 | 1.3863 | 2.8904 |
| LEV | 19,988 | 0.4225 | 0.219 | 0.0084 | 5.9061 |
| Top1 | 19,988 | 0.3226 | 0.147 | 0.0184 | 0.8999 |
| InsInvestorProp | 19,988 | 0.8254 | 0.496 | 0.0000 | 3.0294 |

Regarding our sentiment metrics, we present two related indicators. First, Sentiment is a binary variable derived from the majority vote of our three LLM agents (1 for "Positive", 0 for "Negative"). The mean value of it is 0.9510, which aligns with the general observation that management tends to frame disclosures optimistically, indicating that managers tend to maintain a positive tone regarding the company's business. Our core explanatory variable, avg_sentiment_score, captures the intensity of this sentiment as a continuous measure from 0 to 1, which provides a more nuanced measure than the simple binary classification. The mean of avg_sentiment_score is 0.7967. This further suggests that most listed companies express optimism about their business conditions and future prospects, tending to convey positive and optimistic information to external parties.

### *5.2. Econometric Validation of the MultiSenti Module (Phase 1)*

To examine the association between the sentiment of MD&A with future firm performance and risk, we start our analyses by using Spearman's rank correlation. Table 4 shows the association. It can be observed that the sentiment indicators *Sentiment* and *avg_sentiment_score* are significantly negatively correlated with the indicator of corporate misconduct risk, *Misconduct*. This suggests that disclosures in MD&A provide additional information that is predictive of company performance and risk, which is hard to capture by financial data. Simultaneously, it also provides an initial, unconditional indication that a more negative tone relates to higher violation risk, which supports the fixed-effects tests below.

**Table 4.** Correlation Matrix.

| | Misconduct | Sentiment | avg_sentiment_score | Size | Board | LEV | Top1 | InsInvestorProp |
|---|---|---|---|---|---|---|---|---|
| Misconduct | 1 | | | | | | | |
| Sentiment | −0.079 *** | 1 | | | | | | |
| avg_sentiment_score | −0.113 *** | 0.875 *** | 1 | | | | | |
| Size | −0.054 *** | 0.088 *** | 0.138 *** | 1 | | | | |
| Board | −0.045 *** | 0.015 ** | 0.029 *** | 0.274 *** | 1 | | | |
| LEV | 0.122 *** | −0.053 *** | −0.056 *** | 0.389 *** | 0.105 *** | 1 | | |
| Top1 | −0.115 *** | 0.040 *** | 0.061 *** | 0.190 *** | 0.013 * | −0.017 ** | 1 | |
| InsInvestorProp | −0.080 *** | 0.036 *** | 0.075 *** | 0.452 *** | 0.232 *** | 0.122 *** | 0.491 *** | 1 |

Notes: ***, **, * indicate significance at the 1 percent, 5 percent, and 10 percent levels.

Following our econometric validation protocol as mentioned in Section 4.5, we now present the results of our Phase 1 validation, which is designed to assess the construct validity of the core MultiSenti module. The central question here is whether the sentiment score it generates is a meaningful economic indicator of future corporate misconduct. Table 5 presents the regression results for the impact of sentiment score on the prediction of corporate misconduct. Column 1 represents the logit model result without any fixed effect, while column 2 is the result with the firm fixed effect. In column 3, we present the result that contains the fixed effect with year and firm. There is a difference in observations among models due to the fact that the sample that only has one period would be dropped. We can find that *avg_sentiment_score* has a significantly negative effect on corporate misconduct risk (*Misconduct*). This, once again, demonstrates that the more negative sentiment in the text, the higher the probability of corporate misconduct and the greater the company's risk. The results are consistent with expectations and validate our hypothesis. And they show the value of sentiment information that has been verified in other studies [15,31], as the negative emotions expressed by management in the annual report may reflect the operational difficulties faced by the company, as well as potential issues that management might be attempting to conceal, thus increasing the company's future financial risk and motivation for misconduct. The logit regression results of the models show a similar relationship, significant at the 1% level, further supporting the significant predictive power of negative sentiment in the text for misconduct. Interpreting magnitudes, a one SD increase in the avg_sentiment_score (0.129) reduces the odds of next year's misconduct by about 8–19% (exp[$\beta \times 0.129$], the value decided by $\beta$). Results indicate that the sentiment evaluated by the Large Language Model can present corporate situation and violation risk in efficiency.

**Table 5.** Results.

| | (1) | (2) | (3) |
|---|---|---|---|
| | **Misconduct** | **Misconduct** | **Misconduct** |
| avg_sentiment_score | −1.6202 *** | −0.6629 *** | −0.6467 *** |
| | (−7.4180) | (−2.8199) | (−2.6849) |
| Size | −0.1596 *** | 0.4489 *** | 0.5061 *** |
| | (−3.6736) | (3.6843) | (4.0255) |
| Board | −0.5892 *** | −0.1407 | −0.2702 |
| | (−2.6011) | (−0.4060) | (−0.7716) |
| LEV | 2.0565 *** | 0.1371 | 0.1695 |
| | (10.3073) | (0.5820) | (0.7134) |
| Top1 | −2.7494 *** | 2.4133 *** | 1.9474 ** |
| | (−7.2423) | (3.0143) | (2.2960) |
| InsInvestorProp | −0.2100 * | −0.1024 | −0.1254 |
| | (−1.7398) | (−0.3844) | (−0.4636) |
| Year FE | NO | NO | YES |
| Firm FE | NO | YES | YES |
| N | 19,988 | 4812 | 4812 |

Notes: ***, **, * indicate significance at the 1 percent, 5 percent, and 10 percent levels.

This finding provides strong empirical support for our hypothesis (H1) and, crucially, validates that our MultiSenti module captures a reliable and theoretically consistent signal of misconduct risk.

### *5.3. Robustness Test for Phase 1*

To ensure the robustness of our phase 1 validation findings, we conduct two additional tests: excluding the anomalous COVID-19 period and using an alternative measure for the dependent variable. The special circumstances of the pandemic in 2020 may have affected corporate operations and information transmission. Listed companies may adopt unusually optimistic tones in their Management Discussion and Analysis (MD&A) to cope with the crisis. Consequently, we further exclude the 2020 samples and re-estimate the fixed effect model to verify the robustness of the baseline regression. Table 6 shows that the coefficient for the sentiment score in the annual report is significant at the 1% level, indicating that the sentiment score in the annual report continues to exhibit a robust negative relationship with firm risk after excluding the pandemic's influence. This result is consistent with the baseline regression. Specifically, the coefficient in the test that excludes the 2020 data is larger compared to the baseline regression, suggesting that abnormal changes in management sentiment have a stronger predictive power for firm risk in a normal market environment. The coefficient of *avg_sentiment_score* is −0.8161 and significant at the 1% level. This implies an odds ratio of approximately 0.90 (exp[−0.8161 × 0.129]), suggesting that for a one-SD increase in the sentiment score, the odds of a firm committing a violation decrease by about 10%. This confirms the strong negative association between managerial sentiment and misconduct risk. This further validates the significant role of tone factors in risk early warning in textual analysis, highlighting its predictive power for firm risk.

**Table 6.** Robustness Test.

| | (1) | (2) | (3) |
|---|---|---|---|
| | **Misconduct** | **Misconduct** | **Misconduct** |
| avg_sentiment_score | −1.8408 *** | −0.8044 *** | −0.8161 *** |
| | (−7.6283) | (−2.9409) | (−2.9033) |
| Size | −0.1514 *** | 0.4811 *** | 0.5396 *** |
| | (−3.5501) | (3.6835) | (4.0130) |
| Board | −0.5824 ** | −0.0508 | −0.1965 |
| | (−2.5396) | (−0.1336) | (−0.5114) |
| LEV | 2.1224 *** | 0.2211 | 0.2801 |
| | (10.5648) | (0.7853) | (0.9808) |
| Top1 | −2.5352 *** | 2.4394 *** | 1.8503 ** |
| | (−6.7657) | (2.8776) | (2.0478) |
| InsInvestorProp | −0.2271 * | −0.1259 | −0.1337 |
| | (−1.8972) | (−0.4375) | (−0.4590) |
| Year FE | NO | NO | YES |
| Firm FE | NO | YES | YES |
| N | 16,021 | 3520 | 3520 |

Notes: ***, **, * indicate significance at the 1 percent, 5 percent, and 10 percent levels.

From a model specification perspective, incorporating time fixed effects helps control for macroeconomic fluctuations, such as the impact of changes in the broader environment. Although the sample size decreased due to the exclusion of 2020 data, it still meets the basic requirements for econometric analysis, and the key variables remain significant. This result further strengthens the robustness of the research conclusion, indicating that abnormal changes in the sentiment of management can provide valuable risk signals independent of external shocks.

To enhance the persuasiveness of the results, we estimated the corporate misconduct risks by counting the number of corporate misconduct in the current year. Similar to the baseline model, we fixed individual and year effects in the test. Table 7 shows that

the variable *avg_sentiment_score* yields a consistently significant negative impact on the variable *Violation_count* of enterprises in all models, reinforcing that the framework captures systematic variation in governance risk rather than noise. Collectively, these robustness checks strengthen our confidence in the construct validity of the MultiSenti module's sentiment score.

**Table 7.** Robustness Test (Change dependent variable).

| | (1) | (2) | (3) |
|---|---|---|---|
| | **CMisconduct** | **CMisconduct** | **CMisconduct** |
| avg_sentiment_score | −0.2151 *** | −0.1201 *** | −0.1120 *** |
| | (−9.3258) | (−4.8836) | (−4.5054) |
| Size | −0.0119 *** | 0.0667 *** | 0.1034 *** |
| | (−2.6816) | (6.1383) | (8.5135) |
| Board | −0.0623 *** | −0.0295 | −0.0492 |
| | (−2.7382) | (−0.8811) | (−1.4684) |
| LEV | 0.2648 *** | 0.0382 | 0.0517 |
| | (12.6097) | (1.2149) | (1.6443) |
| Top1 | −0.1980 *** | 0.4435 *** | 0.3230 *** |
| | (−5.2732) | (5.6121) | (3.9790) |
| InsInvestorProp | −0.0168 | −0.0021 | −0.0189 |
| | (−1.3817) | (−0.0849) | (−0.7553) |
| Year FE | NO | NO | YES |
| Firm FE | NO | YES | YES |
| N | 19,988 | 19,988 | 19,988 |

Notes: *** indicate significance at the 1 percent level.

### *5.4. Validation of the MDARisk (Phase 2)*

Having established the construct validity of our core component, we now proceed to the phase 2 validation, which evaluates the practical utility and predictive effectiveness of the entire MDARisk in a realistic forecasting scenario. Table 8 shows the out-of-sample performance.

**Table 8.** Performance of Machine Learning Models with Sentiment Detected by Different Approaches.

| Model | Setting | Recall | Accuracy | F1 | AUC |
|---|---|---|---|---|---|
| GBDT | Baseline Features | **0.6238** | 0.7637 | 0.5667 | 0.6705 |
| | +Traditional Sentiment | 0.6157 | 0.766 | 0.5638 | 0.6692 |
| | +LLM Sentiment (Ours) | 0.5802 | **0.8237** | **0.569** | **0.6728** |
| LDA | Baseline Features | 0.6226 | 0.671 | 0.5213 | 0.6633 |
| | +Traditional Sentiment | 0.6174 | 0.6719 | 0.52 | 0.6639 |
| | +LLM Sentiment (Ours) | **0.6316** | **0.6975** | **0.5375** | **0.6676** |
| Logistic Regression | Baseline Features | 0.6228 | 0.674 | 0.5229 | 0.6639 |
| | +Traditional Sentiment | 0.6204 | 0.6748 | 0.5224 | 0.6641 |
| | +LLM Sentiment (Ours) | **0.6298** | **0.6955** | **0.5359** | **0.6677** |
| SVM | Baseline Features | 0.6182 | 0.6554 | 0.5123 | 0.6622 |
| | +Traditional Sentiment | 0.6163 | 0.6583 | 0.5131 | 0.6613 |
| | +LLM Sentiment (Ours) | **0.6298** | **0.684** | **0.5302** | **0.6674** |
| XGBoost | Baseline Features | **0.5961** | 0.8627 | **0.6002** | 0.7103 |
| | +Traditional Sentiment | 0.5737 | 0.8675 | 0.5823 | 0.7088 |
| | +LLM Sentiment (Ours) | 0.5601 | **0.8767** | 0.5716 | **0.7121** |

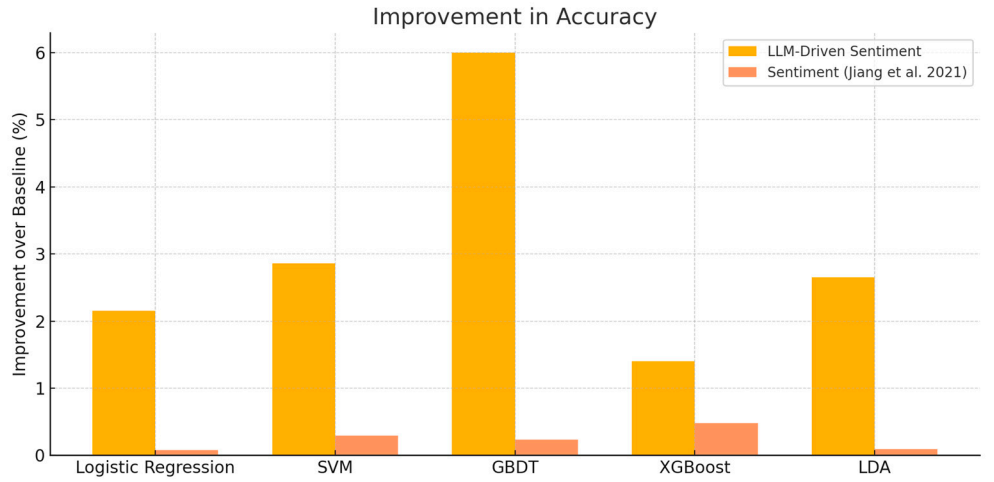
Notes: "Baseline Features" include firm-level controls. "+Sentiment [17]" adds lexicon-based sentiment scores. "+Ours" adds LLM-based sentiment features from the MultiSenti. Bold values in the table indicate the best (highest) performance for each evaluation metric (Recall, Accuracy, F1, AUC) across all model settings.

The superior out-of-sample predictive performance of the MDARisk, particularly when compared to the model augmented with traditional lexicon-based sentiment, provides strong empirical evidence that our MultiSenti module successfully fulfills DR1 (Context-Aware Sentiment Analysis). By consistently outperforming this baseline, our LLM-based approach demonstrates its capability to capture the nuanced, context-sensitive signals within MD&A texts, which translates directly into improved misconduct prediction. Specifically, we can draw two key conclusions. First, compared with the baseline models that use only firm-level features, the inclusion of sentiment scores generated by the MDARisk leads to notable gains across nearly all classifiers, particularly in Accuracy and AUC. For instance, using the GBDT classifier, Accuracy increased from 76.37% to 82.37%, representing a 6 percentage point improvement, while AUC also improved slightly from 0.6705 to 0.6728. Similar trends can be observed in other models. The Logistic Regression model, for example, saw its Accuracy increase from 67.40% to 69.55% and its AUC rise from 0.6639 to 0.6677. The most pronounced gain in Accuracy occurred with the SVM model, which improved from 65.54% to 68.40%, an increase of 2.86 percentage points. This demonstrates that our proposed feature provides significant predictive information beyond traditional firm-level controls.

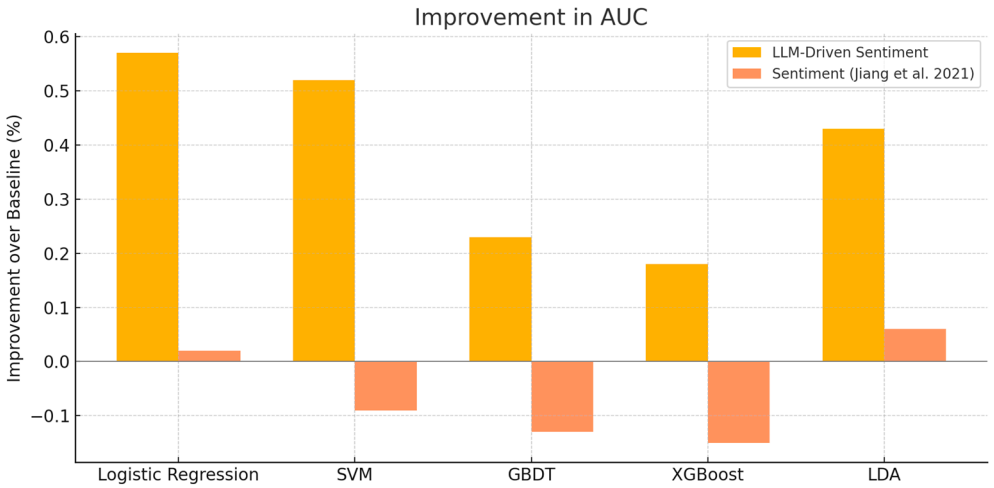
Second, and more importantly, our approach consistently outperforms the traditional lexicon-based sentiment method. The traditional sentiment method based on Jiang et al.'s financial lexicon yields only marginal or inconsistent, and in some cases even negative, improvements in both Accuracy and AUC across most models [17]. In a direct comparison, for every single classifier, the model incorporating our LLM-driven sentiment achieves higher Accuracy and a higher AUC than the model using the traditional sentiment score. This finding robustly validates that MDARisk is more effective at extracting predictively useful signals from MD&A texts than the existing lexicon-based solution.

Figure 6 visualizes the net performance gains over the baseline, offering a clear and compelling summary of our findings. As shown in our visualization, the gains from the traditional method are generally small—often below 1 percentage point—and in some cases even slightly negative. In comparison, MDARisk consistently outperforms the traditional method, yielding noticeable improvements in both metrics. For example, in the GBDT and SVM models, the LLM-based approach improves Accuracy by over 7 and 2.8 percentage points, respectively, and delivers higher AUC values across all classifiers. These results highlight the advantage of leveraging Large Language Models to capture nuanced, context-sensitive sentiment signals from unstructured MD&A texts. Overall, the stark visual evidence further validates the effectiveness of our MDARisk in enhancing out-of-sample prediction of corporate misconduct. The ability of MDARisk to understand context and nuance, a limitation of traditional methods, translates directly into stronger out-of-sample predictive performance.

In summary, the results from this phase successfully validate the MDARisk framework as an effective tool for corporate misconduct prediction. Its ability to understand context and nuance—a key limitation of traditional methods—translates directly into stronger out-of-sample predictive performance, thereby demonstrating its superior practical utility.




(**a**) Improvement in Accuracy




(**b**) Improvement in AUC

**Figure 6.** Comparison analysis of performance improvement. The charts illustrate the superiority of the LLM-Driven Sentiment model over the baseline sentiment analysis method from Jiang et al. [17] across five different classifiers. The top panel shows the percentage improvement in the Area Under the Curve (AUC) metric, while the bottom panel shows the percentage improvement in the Accuracy metric.

## 6. Conclusions

This study undertook a design-science-inspired approach to develop and rigorously evaluate MDARisk, a novel framework for predicting corporate misconduct from Chinese MD&A disclosures. Motivated by the documented limitations of traditional lexicon-based methods, our primary objective was to build an artifact that could extract contextual sentiment with higher accuracy, thereby improving the overall predictive performance for corporate misconduct. Through a comprehensive two-phase evaluation, this research has successfully demonstrated the validity and utility of our proposed solution.

Utilizing a dataset comprising China's A-share listed companies from 2019 to 2023 (MD&A 2018–2022), we first established construct and economic validity. In firm and year fixed effect logit models with a one-period lag and extensive robustness checks, the LLM-derived avg_sentiment_score is significantly negatively associated with next-year misconduct and with the number of violations. This indicates that MultiSenti extracts a sentiment signal that is economically meaningful for governance risk. We then validated system-level predictive utility. Across diverse classifiers and a time-based split, adding the MultiSenti feature improves out-of-sample Accuracy and AUC over both a baseline

without text and a lexicon-augmented baseline, demonstrating incremental information content for forecasting violations.

Through this successful validation, this study makes several key contributions: (i) We propose MultiSenti, a novel Multi-Agent module leveraging LLMs for sentiment analysis, which addresses the inherent limitations of traditional methods in handling challenges such as semantic ambiguity, policy-oriented language nuances, and evolving terminology specific to the Chinese financial context. (ii) We empirically demonstrate the effectiveness of our artifact through a rigorous two-phase validation. Our Phase 1 econometric validation confirmed the construct validity of MultiSenti by establishing a robust negative association between its sentiment score and subsequent corporate misconduct. Furthermore, our Phase 2 predictive validation established the practical utility of the complete MDARisk framework, showing that it significantly enhances out-of-sample prediction accuracy compared to baseline and traditional methods. Furthermore, the application of LLM technology, as implemented in our framework, offers a lightweight and efficient solution for processing vast amounts of textual disclosure data, thereby enhancing regulatory efficiency and supporting more rational assessments of corporate risk based on annual reports.

Based on the research conclusions, we offer practical recommendations from multiple perspectives. First, listed companies should enhance the informational value and transparency of MD&A by providing context-rich, forward-looking narratives on risks and internal controls and by ensuring the content is truthful, accurate, and complete to reduce information asymmetry. Second, the regulatory department can incorporate text-based analytics—particularly LLM-driven sentiment, readability, and uncertainty—into early-warning systems to prioritize reviews of high-risk filings, while strengthening enforcement against false or misleading disclosure. Third, investors should combine sentiment-based text signals with fundamentals to assess ex ante violation risk and exercise caution toward firms with poor disclosure quality. In addition, financial media should standardize the dissemination of information to avoid over-interpreting text features and ensure the effective transmission of market information. Collectively, these measures form a "corporate self-discipline—strengthened regulation—market rationality" governance system, which helps optimize the information environment in capital markets, prevent misconduct risks, and promote the healthy development of markets.

While our findings are robust, we acknowledge certain limitations that present avenues for future research. Our analysis centered on sentiment; future work could integrate other textual dimensions, such as readability or topic modeling, within MDARisk for a more comprehensive risk assessment, which may improve the explanatory power of the model. Beyond expanding the textual scope, the sentiment analysis method itself also has limitations worth noting.

First, the LLM may overestimate the negative emotions because the annual report is replete with a large number of repetitive and standardized legal disclaimers and descriptions of risk factors (for example, "The COVID-19 pandemic may have a significant adverse impact on our future performance"). LLM may judge all these texts as "negative", even if this is just a regular statement required by law rather than the true pessimism of the management. Second, the LLM method may have trouble with discovering antiphrasis or abnormal sentiment, which may be manipulated by the company that has committed misconduct [43]. Additionally, the application of LLMs for sentiment analysis entails substantial computational costs that pose a significant constraint, particularly for large-scale organizational use cases involving high-frequency updates or the processing of massive textual datasets.

Furthermore, the method for improving LLMs' textual analysis for detecting sarcasm, boilerplate legalese, etc., still requires research in the future. Improving the LLM with fine-

tuning and extracting more textual features could significantly boost its adaptability to industry-specific contexts and linguistic subtleties. Such an approach may improve both the practical effectiveness of the LLM-based sentiment analysis across diverse application scenarios.

In sum, we design, build, and validate MDARisk. The framework delivers an economically meaningful measure of managerial tone via MultiSenti and demonstrably improves out-of-sample prediction of corporate misconduct, providing a practical and scalable complement to traditional risk assessment.

**Author Contributions:** Conceptualization, K.Y.; methodology, Y.L. (Yeling Liu); software, Y.L. (Yongkang Liu); validation, K.Y.; formal analysis, Y.L. (Yongkang Liu); investigation, Y.L. (Yeling Liu); data curation, Y.L. (Yongkang Liu); writing—original draft preparation, Y.L. (Yeling Liu); writing—review and editing, K.Y., Y.L. (Yeling Liu) and Y.L. (Yongkang Liu); visualization, Y.L. (Yongkang Liu); supervision, K.Y.; project administration, K.Y.; funding acquisition, K.Y. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by the National Natural Science Foundation of China (Grant Nos. 72401200, 72472103, 72371168), the Shenzhen Stable Support Plan Program for Higher Education Institutions Research Program (No. 20231121164338004), and the High-Level Achievements Cultivation Project of the Third Phase of High-Level University Construction of Shenzhen University (No. 24GSPCG14).

**Data Availability Statement:** The data presented in this study are available on request from the corresponding author. The data cannot be publicly shared as they were purchased from third-party providers, with distribution restricted by licensing agreements.

**Conflicts of Interest:** The authors declare no conflict of interest.

## Abbreviations

The following abbreviations are used in this manuscript:

| | |
|---|---|
| MD&A | Management Discussion and Analysis |
| LLM | Large Language Model |

## Note

[1] Inside this insider trading loophole: What shareholders need to know https://www.businessthink.unsw.edu.au/articles/insider-trading-loophole-shareholders (accessed on 11 September 2025).

## References

1. Baucus, M.S. Pressure, Opportunity and Predisposition: A Multivariate Model of Corporate Illegality. *J. Manag.* **1994**, *20*, 699–721. [CrossRef]
2. Ichev, R. Reported Corporate Misconducts: The Impact on the Financial Markets. *PLoS ONE* **2023**, *18*, e0276637. [CrossRef] [PubMed]
3. Dong, W.; Han, H.; Ke, Y.; Chan, K.C. Social Trust and Corporate Misconduct: Evidence from China. *J. Bus. Ethics* **2018**, *151*, 539–562. [CrossRef]
4. Li, F. Annual Report Readability, Current Earnings, and Earnings Persistence. *J. Account. Econ.* **2008**, *45*, 221–247. [CrossRef]
5. Durnev, A.; Mangen, C. The Spillover Effects of MD&A Disclosures for Real Investment: The Role of Industry Competition. *J. Account. Econ.* **2020**, *70*, 101299. [CrossRef]
6. Loughran, T.; Mcdonald, B. Measuring Readability in Financial Disclosures. *J. Financ.* **2014**, *69*, 1643–1671. [CrossRef]
7. Beracha, E.; Lang, M.; Hausler, J. On the Relationship between Market Sentiment and Commercial Real Estate Performance?A Textual Analysis Examination. *J. Real Estate Res.* **2019**, *41*, 605–638. [CrossRef]
8. Xu, X.; Xiong, F.; An, Z. Using Machine Learning to Predict Corporate Fraud: Evidence Based on the GONE Framework. *J. Bus. Ethics* **2023**, *186*, 137–158. [CrossRef]
9. Wang, R.; Asghari, V.; Hsu, S.-C.; Lee, C.-J.; Chen, J.-H. Detecting Corporate Misconduct through Random Forest in China's Construction Industry. *J. Clean. Prod.* **2020**, *268*, 122266. [CrossRef]
10. Campbell, D.W.; Shang, R. Tone at the Bottom: Measuring Corporate Misconduct Risk from the Text of Employee Reviews. *Manag. Sci.* **2022**, *68*, 7034–7053. [CrossRef]

11. Bel, N.; Bracons, G.; Anderberg, S. Finding Evidence of Fraudster Companies in the CEO's Letter to Shareholders with Sentiment Analysis. *Information* **2021**, *12*, 307. [CrossRef]
12. Henry, E. Are Investors Influenced By How Earnings Press Releases Are Written? *J. Bus. Commun. 1973* **2008**, *45*, 363–407. [CrossRef]
13. Loughran, T.; Mcdonald, B. When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. *J. Financ.* **2011**, *66*, 35–65. [CrossRef]
14. Denecke, K.; Deng, Y. Sentiment Analysis in Medical Settings: New Opportunities and Challenges. *Artif. Intell. Med.* **2015**, *64*, 17–27. [CrossRef]
15. Chen, Y. Forecasting financial distress of listed companies with textual content of the information disclosure: A study based MD&A in Chinese annual reports. *Chin. J. Manag. Sci.* **2019**, *27*, 23–34. [CrossRef]
16. De Kok, T. ChatGPT for Textual Analysis? How to Use Generative LLMs in Accounting Research. *Manag. Sci.* **2025**, *71*, 7223–8095. [CrossRef]
17. Jiang, F.; Meng, L.; Tang, G. Media textual sentiment and Chinese stock return predictability. *China Econ. Q.* **2021**, *21*, 1323–1344. [CrossRef]
18. Velte, P. The Link between Corporate Governance and Corporate Financial Misconduct. A Review of Archival Studies and Implications for Future Research. *Manag. Rev. Q.* **2023**, *73*, 353–411. [CrossRef]
19. Neville, F.; Byron, K.; Post, C.; Ward, A. Board Independence and Corporate Misconduct: A Cross-National Meta-Analysis. *J. Manag.* **2019**, *45*, 2538–2569. [CrossRef]
20. Amiram, D.; Bozanic, Z.; Cox, J.D.; Dupont, Q.; Karpoff, J.M.; Sloan, R. Financial Reporting Fraud and Other Forms of Misconduct: A Multidisciplinary Review of the Literature. *Rev. Acc. Stud.* **2018**, *23*, 732–783. [CrossRef]
21. Lou, Y.-I.; Wang, M.-L. Fraud Risk Factor Of The Fraud Triangle Assessing The Likelihood Of Fraudulent Financial Reporting. *JBER* **2011**, *7*. [CrossRef]
22. Beasley, M.; Carcello, J.; Hermanson, D.; Committee of Sponsoring Organizations of the Treadway Commission. Fraudulent Financial Reporting: 1987–1997: An Analysis of U.S. Public Companies: Research Report. Association Sections, Divisions, Boards, Teams 1999. Available online: https://egrove.olemiss.edu/aicpa_assoc/249/ (accessed on 11 September 2025).
23. Schnake, M.E.; Williams, R.J. Multiple Directorships and Corporate Misconduct: The Moderating Influences of Board Size and Outside Directors. *J. Bus. Strateg.* **2008**, *25*, 1–14. [CrossRef]
24. Aiyesha Dey; Ellen Engel; Xiaohui Liu CEO and Board Chair Roles: To Split or Not to Split? *J. Corp. Financ.* **2011**, *17*, 1595–1618. [CrossRef]
25. Li, J.; Tang, Y. CEO Hubris and Firm Risk Taking in China: The Moderating Role of Managerial Discretion. *AMJ* **2010**, *53*, 45–68. [CrossRef]
26. Meng, Q.; Zou, Y.; Hou, D. Can a short selling mechanism restrain corporate fraud? *Econ. Res.* **2019**, *54*, 89–105.
27. Zou, Y.; Zhang, R.; Meng, Q.; Hou, D. Can stock market liberalization restrain corporate fraud? Evidence from "Shang-hai-Hong Kong stock connect" scheme. *China Soft Sci.* **2019**, 120–134.
28. Cecchini, M.; Aytug, H.; Koehler, G.J.; Pathak, P. Making Words Work: Using Financial Text as a Predictor of Financial Events. *Decis. Support Syst.* **2010**, *50*, 164–175. [CrossRef]
29. Velloor Sivasubramanian, S.; Skillicorn, D. Predicting Fraud in MD&A Sections Using Deep Learning. *J. Bus. Anal.* **2024**, *7*, 197–206. [CrossRef]
30. García, D. Sentiment during Recessions. *J. Financ.* **2013**, *68*, 1267–1300. [CrossRef]
31. Guo, S.; Ning, Q.; Dou, B. Listed Companies' Annual Report Incremental Text Information and Fraud Risk Prediction: From the Perspective of Tone. *Stat. Res.* **2022**, *39*, 69–84. [CrossRef]
32. Loughran, T.; Mcdonald, B. Textual Analysis in Accounting and Finance: A Survey. *J. Account. Res.* **2016**, *54*, 1187–1230. [CrossRef]
33. Berns, J.; Bick, P.; Flugum, R.; Houston, R. Do Changes in MD&A Section Tone Predict Investment Behavior? *Financ. Rev.* **2022**, *57*, 129–153. [CrossRef]
34. Price, S.M.; James, S.D.; David, R.P.; Barbara, A. Bliss Earnings Conference Calls and Stock Returns: The Incremental Informativeness of Textual Tone. *J. Bank. Financ.* **2012**, *36*, 992–1011. [CrossRef]
35. Li, F. The Information Content of Forward-Looking Statements in Corporate Filings—A Naïve Bayesian Machine Learning Approach. *J. Account. Res.* **2010**, *48*, 1049–1102. [CrossRef]
36. Hajek, P.; Henriques, R. Mining Corporate Annual Reports for Intelligent Detection of Financial Statement Fraud—A Comparative Study of Machine Learning Methods. *Knowl. Based Syst.* **2017**, *128*, 139–152. [CrossRef]
37. Huang, A.; Wu, W.; Yu, T. Textual Analysis for China's Financial Markets: A Review and Discussion. *China Financ. Rev. Int.* **2019**, *10*, 1–15. [CrossRef]
38. Bao, Y.; Ke, B.; Li, B.; Yu, Y.J.; Zhang, J. Detecting Accounting Fraud in Publicly Traded, U.S. Firms Using a Machine Learning Approach. *J. Account. Res.* **2020**, *58*, 199–235. [CrossRef]
39. Bello, A.; Ng, S.-C.; Leung, M.-F. A BERT Framework to Sentiment Analysis of Tweets. *Sensors* **2023**, *23*, 506. [CrossRef] [PubMed]

40. Song, P.; Lu, H.; Zhang, Y. Unveiling Tone Manipulation in MD&A: Evidence from ChatGPT Experiments. *Financ. Res. Lett.* **2024**, *67*, 105837. [CrossRef]
41. Xie, D.; Lin, L. Do management tones help to forecast firms' future performance: A textual analysis based on annual earnings communication conferences of listed companies in China. *Account. Res.* **2015**, *2*, 20–27+93.
42. Huang, X.; Teoh, S.H.; Zhang, Y. Tone Management. *Account. Rev.* **2014**, *89*, 1083–1113. [CrossRef]
43. Tan, J.; Wang, X. Corporate Fraud and Manipulation of Annual Report Text Information. *China Soft Sci.* **2022**, *3*, 99–111.
44. Naresh Kumar, K.E.; Uma, V. Intelligent Sentinet-Based Lexicon for Context-Aware Sentiment Analysis: Optimized Neural Network for Sentiment Classification on Social Media. *J. Supercomput.* **2021**, *77*, 12801–12825. [CrossRef]
45. Moradi-Kamali, H.; Rajabi-Ghozlou, M.-H.; Ghazavi, M.; Soltani, A.; Sattarzadeh, A.; Entezari-Maleki, R. Market-Derived Financial Sentiment Analysis: Context-Aware Language Models for Crypto Forecasting. *arXiv* **2025**, arXiv:2502.14897.
46. Araci, D. FinBERT: Financial sentiment analysis with pre-trained language models. In Proceedings of the Eleventh International Conference on Learning Representations (ICLR 2020), Addis Ababa, Ethiopia, 26 April–1 May 2020.
47. Gardazi, N.M.; Daud, A.; Malik, M.K.; Bukhari, A.; Alsahfi, T.; Alshemaimri, B. BERT Applications in Natural Language Processing: A Review. *Artif. Intell. Rev.* **2025**, *58*, 166. [CrossRef]
48. Kim, C.; Zhang, L. Corporate Political Connections and Tax Aggressiveness. *Contemp. Account. Res.* **2016**, *33*, 78–114. [CrossRef]
49. Hevner, A.R.; March, S.T.; Park, J.; Ram, S. Design Science in Information Systems Research. *MIS Q.* **2004**, *28*, 75–105. [CrossRef]
50. Wang, X.; Wei, J.; Schuurmans, D.; Le, Q.; Chi, E.; Narang, S.; Chowdhery, A.; Zhou, D. Self-consistency improves chain of thought reasoning in language models. In Proceedings of the Eleventh International Conference on Learning Representations (ICLR 2023), Kigali, Rwanda, 1–5 May 2023.
51. Dechow, P.M.; Ge, W.; Larson, C.R.; Sloan, R.G. Predicting Material Accounting Misstatements. *Contemp. Account. Res.* **2011**, *28*, 17–82. [CrossRef]
52. Beasley, M.S. An Empirical Analysis of the Relation between the Board of Director Composition and Financial Statement Fraud. *Account. Rev.* **1996**, *71*, 443–465.
53. Guyon, I.; Elisseeff, A. An introduction to variable and feature selection. *J. Mach. Learn. Res.* **2003**, 1157–1182.
54. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic Minority Over-Sampling Technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [CrossRef]
55. Saito, T.; Rehmsmeier, M. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLoS ONE* **2015**, *10*, e0118432. [CrossRef]
56. Davis, J.; Goadrich, M. The Relationship between Precision-Recall and ROC Curves. In *Proceedings of the 23rd international conference on Machine learning, Pittsburgh, PA, USA, 25–29 June 2025*; Association for Computing Machinery: New York, NY, USA, 2006; pp. 233–240.
57. Chung, R.; Firth, M.; Kim, J.-B. Institutional Monitoring and Opportunistic Earnings Management. *J. Corp. Financ.* **2002**, *8*, 29–48. [CrossRef]
58. King, G.; Zeng, L. Logistic Regression in Rare Events Data. *Political Anal.* **2001**, *9*, 137–163. [CrossRef]
59. Nwulu, N.; Oroja, S.; İlkan, M. A Comparative Analysis of Machine Learning Techniques for Credit Scoring. *Information* **2012**, *15*, 4129–4145.


**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.