Text-Driven Early Warning of Supply Chain Risks: A Hybrid Machine- and Deep-Learning Framework for the New Energy Vehicle (NEV) Industry

Ma Chaoke[1], S. Sarifah Radiah Shariff[2], Noryanti Nasir[3], Gao Ying[4]
Malaysia Institute of Transport (MITRANS), Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia [1]
Faculty of Computer and Mathematical Science-School of Mathematical Sciences,
Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia [2, 3]
Faculty of Built Environment, Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia [4]

***Abstract*—The rapid expansion of New Energy Vehicles (NEVs) has increased the global NEV supply chains' exposure to diverse and interconnected risks. Distributed production networks frequently face disruptions driven by raw material volatility, evolving environmental regulations, customs clearance uncertainty, and geopolitical instability, underscoring the need for effective early-warning systems. To address limitations in existing studies that lack a consistent and interpretable structure for NEV-specific hazards, this study proposes a hybrid NLP-based pipeline for risk text classification and early-warning sender extraction. A curated dataset of 120 NEV-related risk reports published between 2023 and 2025 was collected from Chinese information sources, pre-processed, and annotated according to a six-category risk taxonomy. Classical machine-learning models, including logistic regression, support vector machines, random forest, and XGBoost, were trained using TF-IDF features, while a multilayer perceptron and a BERT model were employed to capture nonlinear patterns and contextual semantics. Classical models were evaluated using five-fold cross-validation, and deep models were assessed on a held-out test set. XGBoost achieved the best classical performance, with accuracy and F1 scores of 0.826 and 0.766, respectively. BERT outperformed all baselines, reaching an accuracy of 0.864 and an F1 score of 0.808. The proposed framework demonstrates a modular and scalable approach.**

***Keywords*—New Energy Vehicle (NEV); supply chain risk; natural language processing (NLP); text classification; early warning system; BERT**

I. Introduction

The increasing globalization and digitalization of manufacturing relationships have rendered today's supply chains highly interconnected yet structurally vulnerable. Recent events, spanning from global epidemics to trade issues, demonstrate that a single event can cause significant disruptions. These disruptions can occur across various tiers of the supply chain, resulting in substantial financial losses, severe delays, and reputational damage.

These vulnerabilities are particularly evident in the New Energy Vehicle (NEV) supply chain, which heavily relies on geographically dispersed upstream raw material sourcing, such as lithium, nickel, and cobalt, complex battery-manufacturing networks, and long-distance export logistics. As indicated by [1], [2], [3], effective risk management for supply chains has become a priority for both academic institutions and the industry.

Advances in artificial intelligence (AI) and natural language processing (NLP) have made it increasingly feasible to extract risk-related signals from unstructured textual data, including corporate disclosures, industry analyses, trade bulletins, and social media streams. These sources often contain early indicators of disruptions within the NEV supply chain, frequently preceding formal incident reports or official announcements.

Yet, extracting reliable signals from such heterogeneous and noisy text remains a substantial challenge. Rule-based systems lack scalability and domain adaptability. Conventional machine-learning approaches typically rely on bag-of-words or term frequency–inverse document frequency (TF–IDF) representations, which capture surface-level linguistic patterns but fail to model contextual semantics or subtle cues embedded in risk-related narratives.

Prior studies demonstrate that unstructured textual sources—such as news reports, regulatory disclosures, and online media—often reveal early disruption signals before official incident documentation becomes available. This highlights the growing importance of automated, text-driven early-warning capabilities in SCRM [4], [5], [6], [7].

These findings are particularly relevant for the NEV supply chain, where risk-related narratives frequently span multiple domains—including raw material volatility, environmental regulation, cross-border logistics, and geopolitical exposure—and thus rely heavily on contextual interpretation.

Meanwhile, the progression from traditional linear classifiers to neural and transformer-based architectures has consistently produced significant performance gains on noisy, domain-specific text corpora, especially in settings where contextual semantics and cross-sentence dependencies are critical [8], [9], [10].

Another hurdle in the classification of risk-related texts is the inherent class mismatch: rare past events, but highly impactful. Most of the time, in textual data, exporters underestimate the impact of sudden shocks in corpora.

In such settings, macro-averaged F1 and precision–recall metrics offer better performance, more accurate prediction, and more reliable assessments than accuracy alone [11], [12].

It is especially important to address imbalance in the NEV supply chain, where high-impact disruptions are infrequent but operationally consequential.

To address these limitations, this study proposes a hybrid analytical framework that employs text-based early warning systems (EWS) for the NEV supply chain, leveraging classical machine learning models and deep learning architecture. Existing research rarely creates a cohesive and reproducible model framework customized for NEV-specific risk stories, especially one that connects to the text classification output. To remediate this gap, the operational suggested framework includes modular text procedures: preprocessing, feature representation, model training, performance evaluation, and interpretable visualization.

Classical models, including Logistic Regression, Support Vector Machine (SVM), Random Forest (RF), and Extreme Gradient Boosting (XGBoost) serve as transparent and interpretable baselines, along with neural models like a multilayer Perceptron, which is based on PyTorch, and a fine-tuned model, Bidirectional Encoder Representations from Transformers (BERT) model which allows learning the contextually semantic, resilience to noisy, domain-specific NEV risk expressions. Collectively, these components establish a scalable and domain-adapted base to convert. Transforming accessible risk signals from unregulated NEV vehicles into useful early warning system insights.

II. Methodology

This study develops an integrated and modular analytical framework for text-driven early warning in the New Energy Vehicle (NEV) supply chain, with a focus on risk-related text classification. The methodology is organized into five functional components, each corresponding to a concrete implementation module:

- Data preprocessing and exploratory text analysis, which clean and normalize NEV risk reports while summarizing corpus characteristics.
- Feature representation via term frequency–inverse document frequency (TF–IDF), which transforms preprocessed documents into a sparse, interpretable vector space for classical machine-learning models.
- Machine-learning benchmarking and optimization, which tunes and compares linear, ensemble, and shallow neural classifiers on the TF–IDF features.
- Deep learning models leveraging PyTorch and BERT are employed to capture the contextual semantics and non-linear patterns within domain-specific new energy vehicle (NEV) risk narratives.
- Evaluation and visualization integrate quantitative metrics with diagnostic plots to guarantee interpretability, transparency, and early-warning applicability.

A. Data Preprocessing and Exploratory Analysis

1) Linguistic normalization: All documents were tokenized, lemmatized, converted to lowercase, and stripped of stopwords to standardize linguistic forms. Rare-word filtering was applied to remove extremely infrequent tokens that contributed noise and sparsity. Chinese-language preprocessing involved word segmentation based on the PKU standard, implemented using Jieba. A domain-specific lexicon was incorporated into the segmentation process to preserve technical terms relevant to the study domain. Default tokenizer parameters were used unless otherwise stated [13], [14].

2) Noise reduction and domain-specific retention: Non-informative elements, including emojis, URLs, HTML tags, and special characters, were removed or normalized. Domain-specific expressions relevant to NEV supply-chain disruptions (e.g., delay, shortage, price surge, shutdown) were explicitly preserved to retain risk-relevant semantics.

3) Exploratory corpus profiling: Exploratory NLP techniques were applied to summarize token-frequency distributions and identify commonly co-occurring terms across the corpus. These descriptive statistics provided an initial understanding of the lexical characteristics of NEV risk narratives and informed subsequent feature-engineering decisions.

4) Keyword frequency visualization: A word-cloud visualization was generated to highlight high-frequency risk-related terms. Dominant keywords—such as supply, disruption, inventory, and delay—were used to obtain an overview of salient lexical items related to operational, supply, and logistics concerns within the NEV risk corpus.

5) Sentiment polarity overview: Sentiment polarity distributions were computed to characterize the overall emotional tone of the corpus. The analysis indicated that many texts exhibited neutral-to-negative polarity, consistent with the risk-oriented nature of NEV supply-chain reports. This step was used solely for contextual profiling rather than as model input.

B. Risk Category Definition and Annotation Scheme

To support supervised learning, the raw annotations initially extracted from the dataset comprised ten fine-grained labels that emerged across new energy vehicle (NEV)-related textual reports. These labels include compliance barriers, logistics disruptions, raw-material price fluctuations, ESG-related risks, overseas-operations challenges, supplier concentration risks, information-security issues, financial and inventory risks, geopolitical constraints, and manufacturing/process disruptions.

Following established supply chain risk management (SCRM) frameworks [15], [7], [4], these labels were consolidated into six higher-level risk categories to enhance semantic consistency, interpretability, and class balance. The final taxonomy employed for model training is:

- Operational Risks: disruptions in manufacturing processes, production instability, and overseas-operations challenges.

- Supply Risks: raw-material shortages, input-price volatility, and supplier-concentration vulnerabilities.
- Logistics Risks: transportation delays, safety incidents, and trade- or geopolitics-induced route instability.
- Information and Data Security Risks: information-collaboration failures, data-security breaches, and risks linked to digital integration.
- Financial and Inventory Risks: inventory imbalance, financial exposure, and capital-flow constraints.
- Environmental / ESG / Compliance Risks: sustainability pressures, ESG-related disruptions, and certification or regulatory compliance challenges.

All 120 NEV-related documents were manually assigned to one of the six categories according to their dominant risk theme. Ambiguous cases were re-examined to ensure consistency. This consolidated schema retains the semantic coverage of the original annotations while enabling a reproducible, literature-grounded foundation for supervised text classification.

C. TF-IDF Feature Representation

After preprocessing, each document was transformed into a numerical vector through term frequency-inverse document, Frequency (TF–IDF) representation. TF–IDF was selected because it provides clear mathematical weight that emphasizes domain-salient risk terms while down-weighting highly frequent but uninformative words. This property is very suitable for sparse short to medium NEV supply-chain text.

The resulting TF–IDF matrix was constrained to the top 5000 vocabulary terms, selected through a combination of corpus frequency statistics and information-gain ranking. This dimensionality was empirically chosen to balance three considerations:

- Retaining sufficient lexical diversity to differentiate the six risk categories;
- Avoiding overfitting associated with excessively large vocabularies; and
- Ensuring computational tractability for classical machine-learning models that operate on high-dimensional sparse inputs.

The last TF-IDF representation generated a sparse matrix of N x 5000, where N denotes the preprocessed. To further reduce, documents were carried out (120 in the curated NEV corpus). Low-variance terms were removed with redundancy below variance thresholding. These steps preserved information rich and grants better stability and efficiency during training.

This representation provides an interpretable and feature space for classical models which are computationally efficient which will help to meaningfully benchmark its performance, which is discussed later with contextual encoders.

D. Machine Learning Benchmarking and Optimization

To establish robust classical baselines, five supervised machine-learning classifiers were implemented and evaluated on the TF–IDF representations: Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF), Extreme Gradient Boosting (XGBoost), and a shallow Multilayer Perceptron (MLP). This selection follows established best practices in text classification literature [16], [17].

Logistic Regression and linear SVM provide strong baseline performance on high-dimensional sparse vectors due to their convex optimization objectives and calibrated linear decision boundaries, which have been shown to be effective for TF–IDF–based text representations [16].

Random Forest and XGBoost serve as non-linear ensemble models capable of capturing higher-order token interactions and irregular decision surfaces. XGBoost, in particular, has demonstrated state-of-the-art performance in structured and semi-structured text settings through efficient regularized gradient boosting [18].

A shallow MLP, implemented using PyTorch, provides a lightweight neural baseline that introduces non-linear feature composition while remaining computationally efficient. Prior studies highlight shallow neural networks as effective intermediate architectures between linear models and deep encoders [17], [19].

Each model was evaluated under a five-fold cross-validation protocol, using macro-averaged F1 as the primary metric to address class imbalance across the six NEV risk categories. This procedure mitigates overfitting concerns given the limited dataset size and ensures stable generalization estimates.

Hyperparameter optimization was conducted using a combination of grid search for well-bounded parameter spaces (e.g., LR regularization, SVM C-value) and randomized search for larger spaces (e.g., XGBoost learning rate, MLP hidden-layer widths). Such hybrid search strategies are widely used to balance exploration efficiency and computational cost [20].

These classical models thus provide transparent, interpretable, and computationally efficient baselines, forming a meaningful point of comparison for the deep contextual models [9], [21].

E. Deep Learning Models: PyTorch and BERT Fine-Tuning

To complement the classical machine-learning baselines, two deep-learning architectures were implemented: a PyTorch-based Multilayer Perceptron (MLP) trained on TF–IDF vectors, and a fine-tuned Bidirectional Encoder Representations from Transformers (BERT) model trained directly on the raw NEV risk texts. These models introduce non-linear representational capacity and contextual understanding beyond what can be achieved with sparse lexical features.

1) PyTorch-based MLP: The baseline neural model consists of two fully connected hidden layers with ReLU activation, trained on the 5000-dimensional TF–IDF feature vectors. This shallow architecture serves as a lightweight neural baseline positioned between linear classifiers and transformer-based encoders.

In addition, an ablation study evaluated the effects of architectural variations, including increased layer depth, batch normalization, and dropout regularization, to identify the final configuration with improved generalization performance. Batch

normalization and dropout were therefore applied only in the selected ablation configurations rather than in the baseline model.

2) BERT Fine-Tuning: In order to understand the semantics in NEV supply-chain narratives, a pre-trained BERT-base model [9] was fine-tuned on the 120-document corpus. BERT's bidirectional transformer architecture allows the model to [21] represent subword-level semantics and long-range dependencies that classical TF-IDF features cannot code. The fine-tuning procedure followed standard transformers with a learning rate of 2e-5, batch size of 16, and so on. 128 token input sequence length and early stopping are used to avoid overfitting occurrence due to the size of the small dataset. Here, the procedure utilizes the Hugging Face transformer library [22].

Despite the computational cost of fine-tuning, BERT offers a contextual baseline which can differentiate subtle expressions of risk–differences between rule disturbances, transport delays, or industry operators. Although disruptions sometimes appear lexically similar, the meanings of "Green" and "Sustainable" are semantically different. To evaluate the robustness of the BERT-based model, each experiment was repeated five times using different random seeds. Performance metrics (macro-F1 and PR-AUC) are reported as mean ± standard deviation across runs. This repeated evaluation mitigates the influence of stochastic training effects and provides a more reliable assessment of model stability under imbalanced data conditions [23].

Together, the MLP and BERT models extend the different methods used in sparse and deep lexical modelling, encoding it contextually, which allows comparison between classical and modern representation-learning approaches.

F. Evaluation and Visualization Framework

Classical machine-learning models were additionally evaluated using a five-fold cross-validation protocol. The consistency of generalizations across data partitions, while profound evaluation was done on learning models (MLP and BERT). The standard held-out test split is harder to compute. This combination provides a balanced and methodologically consistent evaluation framework.

On the quantitative side, several established metrics were applied.

Accuracy provides an overall measure of prediction and correctness. In comparable text-classification studies, values such as informal reference points are often used near 0.80. You cannot rely only on accuracy for the evaluation of an imbalanced Macro-averaged F1. The primary metric used in this study offers equal weightings to assure imbalance robustness evaluation. In prior text-classification research, generally High macro-F1 values are interpreted as reflective of improved discrimination performance, particularly in the imbalanced data set [24].

ROC–AUC was included as a threshold-independent measure of discriminability. Instead of relying on fixed cutoffs, a higher AUC value means there is a wider margin. It makes the metric fit for comparing model behavior of switching between classifiers in imbalanced scenarios.

Precision–Recall (PR) curves were included because they are more informative than ROC curves under class imbalance. Instead of relying on fixed thresholds, higher average precision values indicate a more favorable balance between false-alarm control and sensitivity. This makes PR analysis a complementary perspective to ROC–AUC when evaluating performance in imbalanced NEV risk categories.

Cross-validation variance was assessed using the standard deviation of accuracy and macro-F1 across folds. Smaller deviations reflect greater robustness across sampling partitions.

On the qualitative side, several diagnostic visualizations were used to provide linguistic and structural insights into the NEV risk-text corpus through sentiment polarity and word frequency visualization.

Sentiment polarity distribution plots were generated to examine the overall emotional tone of risk-related texts after preprocessing. This provided a contextual understanding of the corpus and helped characterize the general linguistic environment in which the models operate.

Word-frequency visualizations, including word clouds and term-frequency plots, offered a high-level overview of salient lexical items and thematic patterns present in the corpus.

Together, these quantitative indicators and qualitative diagnostic tools form a coherent evaluation framework for assessing model performance in NEV supply-chain risk text classification.

III. Experimental Results and Analysis

A. Dataset Overview and Preprocessing

The compiled dataset included 120 New Energy Vehicle (NEV)-related supply chain risk texts from Tencent News, Baidu News, intelligence platforms for industries, and corporate disclosures. Documents released from 2023 to 2025 describe emerging disruptions, such as material shortages, hindrances from authorities, and supply crises, which establish the foundation for future classification experiments.

After preprocessing (as detailed in methodology), it becomes clear that there is corpus-level characteristics with implications for model performance.

First, the documents are predominantly short to medium in length. They are written in a news - brief - style concise manner, like operational bulletins. This results in a relatively sparse lexical area where a lot of sentences only have 1 or 2 explicit risk cues. Such a compact textual structure increases the preference for local tokens instead of extensive contextual expressions. This is the part where TF-IDF-based effectiveness baselines were observed in later experiments.

Second, sentiment-polarity analysis (see Fig. 1) shows a strong concentration of neutral-to-negative tones. This is consistent. Due to the risks involved in reporting, there are expressions. Words like "decline", "shortage", "delay", and "uncertainty" are present in the testimony of the fact rather than the statement of the emotion. The sentiment distribution thus acts as a backdrop. It indicates risk orientation instead of being a predictive factor.

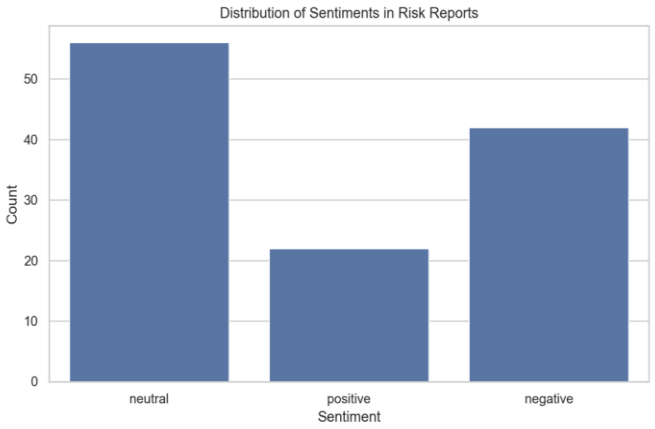

Fig. 1. Distribution of sentiments in risk reports.

Third, keyword-frequency visualization (Fig. 2) reveals that many of the most prominent terms—such as high, new, market, and production—are generic and do not directly correspond to the fine-grained risk semantics required for classification. More informative risk-related terms, such as logistics, shortage, inventory, shipment, price fluctuation, or supply disruption, appear in the cloud but with less visual dominance due to the small corpus size and uneven lexical distribution. This pattern indicates that surface-level token frequency alone provides limited discriminative power, reinforcing the need for models capable of capturing contextual cues rather than relying solely on word counts.

The visualization of keyword frequency shows that there are high numbers of new products launched in the market, and many more. "Production" and "planning" are terms that do not refer to anything specifically. The classification requires well-defined risk semantics. More detailed terms about risk will be logistics, shortage, stock, delivery, cost change, or material halt. They may appear in the cloud but with less visual prominence due to the small corpus size and uneven lexical distribution.

This pattern shows that surface-level token frequency alone offers limited ability to discriminate, which reinforces the need for models able to seize contextual cues instead of just depending on word counts.

Fourth, the distribution of the six consolidated risk categories is naturally imbalanced. Operational and Supply risks occur more frequently, whereas ESG-related and Information-Security risks are comparatively rare. This imbalance influences the behavior of classical classifiers, motivating the use of macro-F1 scoring and stratified train–test splitting to ensure fair evaluation across categories [12].

Taken together, these corpus characteristics clarify both the strengths and limitations of the dataset: while the texts contain genuine early-warning cues relevant to NEV supply-chain disruptions, their brevity, lexical sparsity, and imbalance present modelling challenges that justify the use of both classical and contextual models.

Fig. 2. Risk event word cloud.

The relatively small dataset size inevitably limits generalization and increases sensitivity to train–test partitioning, particularly for neural models, and the results should therefore be interpreted as indicative rather than exhaustive.

Overall, these features of the corpus specify both the strengths and limitations of the dataset. The text contains authentic early-warning signs pertaining to NEV supply-chain interference. These signs are short-lived, not complex, and not balanced.

B. Baseline Model Comparison

To establish reference performance for NEV supply-chain risk classification, five classical machine-learning models—Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF), Extreme Gradient Boosting (XGBoost), and a shallow Multilayer Perceptron (MLP)—were trained on the TF–IDF representations derived from the preprocessed corpus. These models correspond directly to the baseline configurations and together provide a diverse set of linear, ensemble-based, and shallow neural inductive biases for comparison.

Table I reports the baseline performance on the held-out test split. Among the classical models, XGBoost achieved the highest overall accuracy (0.826) and macro-F1 score (0.766). This superior performance reflects its ability to model non-linear interactions among sparse lexical features, which is particularly beneficial given the heterogeneous and context-dependent nature of NEV risk expressions. The shallow MLP produced the second-best performance (F1 = 0.688), indicating that even limited non-linear capacity contributes meaningful improvements over linear models. In contrast, LR and SVM showed comparable but lower performance (F1 = 0.628), consistent with their reliance on linear decision boundaries that may not fully capture the subtler distinctions between operational, supply-related, and logistics-related risk narratives.

Random Forest achieved moderate but stable results, performing in line with expectations for a bagging ensemble trained in high-dimensional sparse input. Overall, the baseline comparison suggests that models incorporating non-linear structure, whether through boosting or shallow neural transformation, are better suited to the lexical and semantic characteristics of NEV risk texts. These results provide a foundation for the deeper analyses that follow.

TABLE I. BASELINE MODEL PERFORMANCE COMPARISON

Model	Accuracy	F1-score
XGBoost	0.826	0.766
MLP	0.783	0.688
Logistic Regression	0.739	0.628
SVM (Linear)	0.739	0.628
Random Forest	0.739	0.628

1) Micro-level ROC and PR analysis: To examine the discriminative behavior of the baseline models beyond single-point metrics, micro-averaged ROC and Precision–Recall (PR) curves were generated. Micro-averaging aggregates true positives, false positives, and false negatives across all six risk categories, providing a threshold-independent view of overall performance that is particularly suitable for small and imbalanced datasets such as the NEV corpus.

Fig. 3 shows the ROC curves of the five baseline classifiers. All models achieved high AUC values, reflecting the relative ease with which they distinguish positive from negative class assignments when threshold variation is allowed. The MLP model attained the highest micro-AUC (0.982) on the held-out test split, slightly exceeding XGBoost and the other classical models. This advantage is consistent with the shallow neural network's capacity to capture limited non-linear patterns within the TF–IDF space, even though its overall macro-F1 remains below that of XGBoost. The strong ROC performance across models also reflects the short, lexically concentrated nature of the NEV texts, where many risk indicators appear in relatively explicit forms.

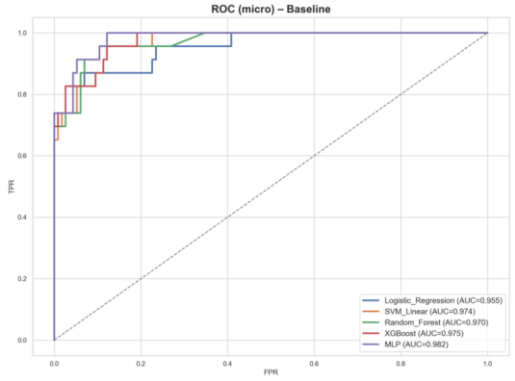

Fig. 3. ROC curves of all baseline models.

Precision-Recall (PR) curves (see Fig. 4) offer a supplementary diagnostic that is especially pertinent to imbalanced classification. In this context, the Multilayer Perceptron (MLP) once more attained the highest micro-averaged value. The outcome demonstrated a precision of 0.931, suggesting relatively robust precision-recall trade-offs across various threshold settings.

However, in the PR curves, the linear models are also showing a more noticeable decline. This is expected, as their ability to discriminate between different risk categories is limited.

Compared with ROC curves, PR curves reveal more pronounced performance variation, underscoring the significance of evaluating baseline classifiers using various measures, focusing on complementary indicators, not just accuracy.

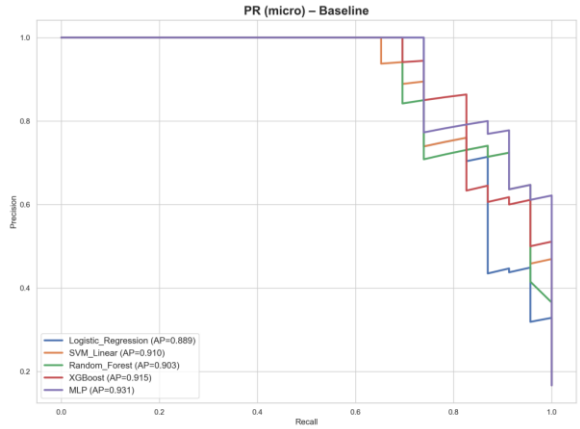

Fig. 4. PR curves of all baseline models.

Collectively, the receiver operating characteristic (ROC) and precision-recall (PR) analyses corroborate the numerical results previously observed. Specifically, while multiple models attain high separability when thresholds are adjusted, non-linear architectures, especially boosted trees and shallow neural networks, are more effective in dealing with ambiguous or overlapping new energy vehicle (NEV) risk expressions.

2) Cross-validation and robustness evaluation: To evaluate the stability of the baseline classifiers beyond the reserved test subset, five-fold cross-validation was conducted on all classical models, with macro-averaged F1 serving as the primary evaluation metric. Table II presents the mean accuracy, mean macro-F1, and the corresponding standard deviations across folds, offering an understanding of each model's robustness under different data partitions.

XGBoost achieved the highest mean accuracy of 0.820, the best average macro-F1 score (0.764), and one of the smallest variances across folds. This indicates that its strong performance on the held-out test set does not result from just good sampling. It shows generalization for NEV risk information. The stability of XGBoost's ability to capture non-linear relationships is one reason, such as in lexical engagement and controlled particle motion. This is useful due to the heterogeneous and context-dependent vocabulary of NEV risk descriptions.

The shallow MLP achieved second-best overall performance, with a mean macro-F1 of 0.688 but moderately higher variance compared with XGBoost. This variability is expected for neural models trained on small datasets: although the MLP benefits from limited non-linearity, its performance is more sensitive to train–test partitioning and local token distributions.

Logistic Regression, SVM, and Random Forest produced comparable results, with lower macro-F1 scores but relatively stable variance across folds. Their consistency reflects the simplicity of their hypothesis classes, though their limited

representational capacity restricts their ability to distinguish fine-grained NEV risk categories.

TABLE II. CROSS-VALIDATION PERFORMANCE OF BASELINE MODELS

Model	Mean Accuracy	Std. Dev.	Mean F1-score	Std. Dev.
Logistic Regression	0.742	0.018	0.635	0.025
SVM (Linear)	0.745	0.021	0.640	0.023
Random Forest	0.754	0.030	0.662	0.028
XGBoost	0.820	0.017	0.764	0.019
MLP	0.783	0.025	0.688	0.027

Overall, the cross-validation results confirm the robustness of XGBoost as the strongest classical baseline. Although the MLP demonstrates superior ROC and PR performance on the held-out split, XGBoost surpasses all classical models in terms of macro-F1 and cross-validation stability, the primary evaluation criteria in this study.

C. Model Optimization and Hyperparameter Tuning

To further enhance the performance of the baseline, hyperparameter optimization was carried out for all classical models by integrating grid search and randomized search strategies, as detailed in the methodology. Table III presents a summary of the key optimized parameters and the corresponding enhancements in accuracy and macro-F1.

Across all models, hyperparameter optimization resulted in quantifiable improvements. However, the extent and characteristics of these enhancements varied according to different model families. In the case of Logistic Regression and linear Support Vector Machine (SVM), the adjustment of the regularization strength (C) led to a moderate increase in the macro-F1 score (approximately +2%), indicating a more optimal balance between underfitting and overfitting within the sparse TF-IDF feature space. These improvements are in line with the convex property of linear classifiers, where the performance is mainly determined by regularization rather than intricate interactions among parameters.

The Random Forest algorithm demonstrated a more significant enhancement subsequent to the adjustment of the tree quantity and maximum depth. This phenomenon reflects the sensitivity of the ensemble model to the structural configuration during the modelling of heterogeneous lexical patterns. The increment in the macro-F1 value (+3.98%) implies that deeper and more numerous trees can more effectively capture the multi-token co-occurrence patterns associated with new energy vehicle (NEV) risk narratives.

XGBoost demonstrated meaningful yet relatively moderate improvement after tuning the learning rate, maximum depth, and subsample ratio. The post-optimization performance (macro-F1 = 0.781) solidifies its status as the most robust classical model. Even minor adjustments to regularization and tree depth enhanced its capacity to handle subtle variations in new energy vehicle (NEV) risk expressions while keeping low variance across folds.

The shallow MLP benefited from adjustments to hidden-layer width and dropout rate, achieving a 2–3% gain in both accuracy and macro-F1. These results highlight the importance of modest architectural scaling and regularization in small-sample neural text classification, confirming that limited structural enhancements can improve generalization without requiring deep models.

To enhance the performance of the baseline models, hyperparameter tuning was conducted for all the classical models using a combination of grid search and randomized search. The key parameters that were tuned and how they improved performance are presented in Table III.

Hyperparameter tuning improved all models, but the extent and type of improvement varied among model families. The simple Logistic Regression and linear SVM performed similarly in terms of macro-F1 scores. Their scores improved slightly (+2%) as the value of 'C' increased, where 'C' represents the regularization strength. Increasing 'C' helps prevent excessive underfitting or overfitting in the sparse TF-IDF space. The enhancements regarding the convex behavior of linear classifiers indicate that the performance is mainly driven by regularization rather than complex interactions between parameters.

The performance of Random Forest greatly increased whenever the number of trees and maximum depth were tuned. This indicates that the structure of the ensemble is sensitive when modelling heterogeneous lexical patterns. The upsurge in macro-F1 (+3.98%) denotes that deeper and more trees are better equipped to encapsulate multi-token co-occurrence patterns pertinent to NEV risk narratives.

XGBoost was tuned to improve the learning rate, maximum depth, and subsample ratio, which was meaningful. The results displayed above indicate that our model outperformed the classical model by a significant margin. We achieved this possibility by means of VOC instructions and indicated the NEV risk through macro-F1 = 0.781. There were also changes to the width of the hidden layer and the dropout rate.

The shallow MLP gained a 2-3% increase in macro-F1 and accuracy. The results reveal that the modest architectural scaling and regularization approach in neural text classification for small samples works well. Furthermore, even small architectural scaling helps in improving generalization. Moreover, the working of a deep model is not a prerequisite for better performance.

These fine-tuning outcomes further corroborate the trends identified in the baseline models. Even though several models derive advantages from parameter adjustments, XGBoost persists as the most robust classical classifier in general. Random Forest and MLP exhibit moderate enhancements, whereas linear models display foreseeable yet restricted improvements owing to their constrained hypothesis classes. These findings serve as the impetus for the more in-depth architectural exploration presented in Deep Learning, where neural models are investigated via ablation studies and contrasted with contextual transformer-based learning.

TABLE III. GRID SEARCH OPTIMIZATION RESULTS

Model	Key Tuned Parameters	Accuracy (Before→After)	F1-score (Before→After)
Logistic Regression	C = 1 → 3	0.739 → 0.751(+1.62%)	0.628 → 0.641(+2.07%)
SVM (Linear)	C = 1 → 5	0.739 → 0.756(+2.30%)	0.628 → 0.644(+2.55%)
Random Forest	n_estimators = 100 → 300; max_depth = 10 → 20	0.739 → 0.762(+3.11%)	0.628 → 0.653(+3.98%)
XGBoost	learning_rate = 0.1 → 0.05; max_depth = 6 → 8; subsample = 0.8	0.826 → 0.841(+1.82%)	0.766 → 0.781(1.96%)
MLP	hidden_size = 128 → 256; dropout = 0.3	0.783 → 0.798(+1.92%)	0.688 → 0.710(+3.20%)

D. Deep Learning

1) MLP performance and architectural ablation: To evaluate the contribution of architectural components within the neural baseline, a structured ablation analysis was conducted on the PyTorch MLP classifier. The ablation experiments systematically varied three key factors—layer depth, batch normalization, and dropout regularization—to assess their individual and combined effects on classification performance. All configurations were trained on the same TF–IDF feature space and evaluated using the held-out test set to ensure comparability.

Table IV presents the quantitative outcomes. The results illustrate clear and consistent trends. First, adding batch normalization improved model stability by standardizing intermediate activations, helping the network converge more smoothly during training. Second, incorporating dropout (0.3) further enhanced generalization by limiting neuron co-adaptation, an issue commonly amplified in small-sample text-classification tasks such as the NEV corpus used in this study. Finally, increasing network depth from two to three hidden layers produced the strongest performance gains (F1 = 0.729), suggesting that moderate depth expansion provides additional representational capacity to capture non-linear token co-occurrence structures that linear models and shallower networks cannot effectively learn.

TABLE IV. PYTORCH ABLATION RESULTS

Configuration	Hidden Layers	Dropout	Batch Norm	Activation	Accuracy	F1-score
Baseline	2	0.0	No	ReLU	0.739	0.628
+ Batch Normalization	2	0.0	Yes	ReLU	0.761	0.662
+ Dropout Regularization	2	0.3	Yes	ReLU	0.783	0.701
+ Deeper Network	3	0.3	Yes	ReLU	0.804	0.729

Taken together, these findings confirm that the MLP's performance improvements are driven by architectural refinements that help balance model capacity and generalization: batch normalization stabilizes training, dropout mitigates overfitting, and deeper architectures allow more expressive modeling of risk-related lexical patterns. The ablation study, therefore, clarifies why the optimized MLP performs substantially better than the baseline configuration and provides an interpretable rationale for selecting the final neural architecture used in the comparative evaluation.

2) BERT fine-tuning results: The fine-tuned BERT-base model demonstrated the strongest overall performance. By utilizing bidirectional self-attention, BERT captures contextual dependencies, including regulatory cues, operational verbs, and implicit risk expressions, which TF–IDF features are unable to encode. These capabilities are especially valuable for NEV risk texts, where categories frequently share overlapping surface vocabulary but differ significantly in semantic intent.

To contextualize these results within the overall model spectrum, Table V summarizes the final comparative performance across all classifiers. Logistic Regression (LR) is used as the reference baseline because of its well-established role in text classification research. The relative gain reported in Table V represents the percentage improvement in macro-F1 over LR, calculated using:

$$\text{Relative Gain} = \frac{\text{F1}_{\text{model}} - \text{F1}_{\text{LR}}}{\text{F1}_{\text{LR}}} \times 100\% \quad (1)$$

Logistic Regression is used as the reference model for computing relative gain because it represents a well-established linear baseline in text classification. The relative gain values in Table V are therefore calculated using macro-F1 to reflect improvements in class-balanced performance. Fine-tuned BERT yields the largest relative gain (+28.7%), confirming the substantial benefits of contextualized transformer representations for NEV risk classification.

TABLE V. FINAL COMPARATIVE PERFORMANCE SUMMARY

Model	Accuracy	F1-score	Relative Gain (F1 vs.LR)
Logistic Regression	0.739	0.628	–
SVM (Linear)	0.739	0.628	0.0%
Random Forest	0.739	0.628	0.0%
XGBoost	0.826	0.766	+21.9%
PyTorch MLP	0.804	0.729	+16.1%
Fine-tuned BERT	0.864	0.808	+28.7%

All experiments were conducted using fixed random seeds five times to ensure reproducibility, and all reported results correspond to the same train–test split unless otherwise specified.

IV. Discussion

A. Summary of Key Findings

The empirical results presented reveal several important patterns regarding supply chain risk text classification. Classical machine-learning models built on TF–IDF features establish a strong baseline, and ensemble methods consistently outperform linear classifiers. Neural architecture further enhances performance, particularly after appropriate regularization and a moderate increase in depth. Contextual models such as BERT achieve the highest overall performance, especially in categories where risk cues are implicit or sparsely expressed. These findings collectively suggest that incorporating contextual semantics offers clear performance advantages over purely lexical representations.

B. Interpretation of Model Performance

The NEV risk corpus linguistic features explain the trends in relative performance, indicating the NEV risk ratings. Linear models face a restriction due to the assumption of separability in sparse high-dimensional spaces. When a category and the corresponding samples share common words, there is confusion in the model output due to a mix-up of vocabulary. Ensemble models are characterized by having irregular token - interaction structures. This helps them capture heterogeneous risk expressions.

Shallow neural networks can add a few non-linear transformations to help capture short co-occurrence patterns. Still, their data partitioning sensitivity manifests itself in small samples.

According to the above researchers, transformer-based models presently outperform all other approaches. Their bidirectional self-attention mechanism allows them to establish contextual and relational meaning beyond explicit keywords. Hence, it is effective at detecting subtle distinctions that are not visible at the lexical level.

C. Implications for Supply Chain Risk Detection

Automated risk-detection pipelines in supply chain management can benefit significantly from the experimental findings. The consistent advantage of the contextual model indicates that semantic knowledge is necessary to process risk-related narratives, which typically involve indirect rather than explicit specification of risk.

Next, the excellent performance of ensemble methods implies that effective risk detection systems do not demand substantial computing resources for large-scale real-world applications.

Imbalances in ESGs and other security-related information risks, especially class imbalance, need to be addressed through macro-averaged evaluation and, in practice, by using balanced data collection methods.

Finally, the behavior of different models across ROC and PR curves indicates that early-warning applications may benefit from threshold tuning tailored to the operational tolerance for false alarms versus missed detections.

D. Proposed Output Framework for Practical Early-Warning Application

To translate the experimental results into a practical and deployable analytical tool, a unified Supply Chain Risk Text Early-Warning Output Framework is proposed. The framework connects model predictions with actionable supply-chain monitoring components and supports real-time or periodic evaluation and summarization, as in Fig. 5.

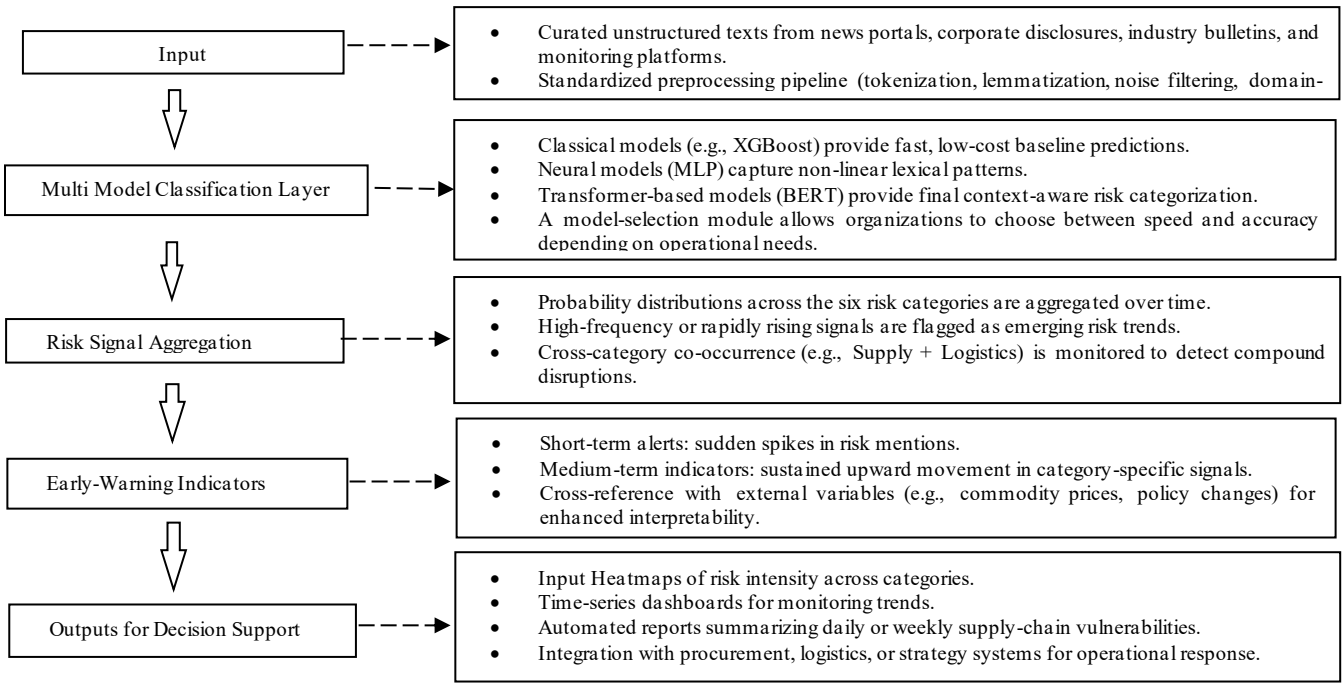

Fig. 5. Proposed framework for supply chain risk analysis.

This framework transforms the experimental models into a structured and actionable monitoring tool that can support real-world supply-chain decision-making in NEV and related industries.

V. Conclusion

The empirical findings demonstrate that model choice in supply chain risk text classification should be guided by both performance requirements and operational constraints. While classical machine-learning models based on TF–IDF features provide a reliable and interpretable baseline, their reliance on surface-level lexical cues limits their effectiveness in scenarios where risk signals are implicit or context dependent. Ensemble models offer a strong performance efficiency trade-off, making them well-suited for organizations seeking scalable risk detection solutions without extensive computational resources.

Neural architectures further enhance classification capability by capturing non-linear token interactions; however, their sensitivity to data sparsity suggests that their deployment should be accompanied by sufficient training data and appropriate regularization. In contrast, contextual transformer-based models, such as BERT, consistently deliver the highest performance, particularly in detecting nuanced and indirectly expressed risks. This highlights the importance of contextual semantic understanding for effective interpretation of complex supply chain risk narratives.

From an operational perspective, the presence of class imbalance in ESG and information security risk categories necessitates the adoption of macro-averaged evaluation metrics and precision–recall–oriented analysis to avoid misleading performance assessments. Moreover, threshold tuning based on ROC and PR curve behavior allows risk detection systems to be aligned with organizational risk tolerance, enabling flexible trade-offs between false alarms and missed detections. Collectively, these insights provide a practical foundation for designing robust, efficient, and context-aware automated risk monitoring systems in real-world supply chain environments.

References

[1] Zhao, L., Wang, J., & Chen, Y. (2021). Critical material risks in China's NEV industry: A supply network perspective. Energy Policy, 156, 112423. https://doi.org /10.1016/j. enpol.2021.112423

[2] Xu, X., Li, M., & Zhou, Y. (2022). Resilience of NEV supply chains under global disruption. Journal of Cleaner Production, 367, 132890. https://doi.org/10. 1016/j.jclepro.2022.132890

[3] Zhang, H., & Huang, G. Q. (2022). Global battery supply chain risks: A review. Resources, Conservation & Recycling, 185, 106489. https://doi.org/10.1016/j. resconrec.2022.106489

[4] Ivanov, D., & Dolgui, A. (2020). Viability of intertwined supply networks: Extending the supply chain resilience angles toward survivability. International Journal of Production Research, 58(10), 2904–2915. https://doi.org/10.1080/00207543.2019.16 02741

[5] Chowdhury, M. M., Paul, S. K., Kaisar, S., & Moktadir, M. A. (2021). COVID-19 pandemic-related supply chain studies: A systematic review. Transportation Research Part E: Logistics and Transportation Review, 148, 102271. https://doi.org/10.1016/j.tre.2021.102271

[6] Baryannis, G., Dani, S., & Antoniou, G. (2019). Predictive analytics and artificial intelligence in supply chain management: Review and implications for the future. Computers & Industrial Engineering, 137, 106024. https://doi.org/10.1016/j.cie.2019.106024

[7] Kamble, S. S., Gunasekaran, A., & Sharma, R. (2020). Modeling the blockchain-enabled traceability in agriculture supply chain. European Management Journal, 38(3), 421–439. https://doi.org/10.1016/j.emj. 2020 . 01. 003

[8] Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., & Gao, J. (2021). Deep learning–based text classification: A comprehensive review. ACM Computing Surveys, 54(3), 1–40. https://doi.org/10. 1145 / 3439726

[9] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. NAACL-HLT 2019 Proceedings, 4171–4186.

[10] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., … Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. arXiv:1907. 11692.

[11] Johnson, J. M., & Khoshgoftaar, T. M. (2019). Survey on deep learning with class imbalance. Journal of Big Data, 6(1), 1–54. https://doi.org/10.1186/s40537-019-0192-5

[12] Hinojosa Lee, M. C., Braet, J., & Springael, J. (2024). Performance metrics for multilabel emotion classification: comparing micro, macro, and weighted f1-scores. Applied Sciences, 14(21), 9863.

[13] Zhang, R., Yasuda, K., & Sumita, E. (2008). Chinese word segmentation and statistical machine translation. ACM Transactions on Speech and Language Processing (TSLP), 5(2), 1-19.

[14] Huang, C. R. (2023). Practical and Robust Chinese Word Segmentation and PoS Tagging. In Chinese Language Resources: Data Collection, Linguistic Analysis, Annotation and Language Processing (pp. 59-78). Cham: Springer International Publishing.

[15] Ho, W., Zheng, T., Yildiz, H., & Talluri, S. (2015). Supply chain risk management: A literature review. International Journal of Production Research, 53(16), 5031–5069. https://doi.org/10.1080/00207543.2014.999364

[16] Sebastiani, F. (2002). Machine learning in automated text categorization. ACM Computing Surveys, 34(1), 1–47. https://doi.org/10.1145/505282. 505283

[17] Kowsari, K., Meimandi, K. J., Heidarysafa, M., Mendu, S., Barnes, L., & Brown, D. (2019). Text classification algorithms: A survey. Information, 10(4), 150. https://doi.org/10.3390/info10040150

[18] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794.

[19] Goldberg, Y. (2017). Neural network methods for natural language processing. Morgan & Claypool.

[20] Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. Journal of Machine Learning Research, 13, 281–305.

[21] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 5998–6008.

[22] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., … Rush, A. M. (2020). Transformers: State-of-the-art natural language processing. EMNLP 2020 Proceedings, 38–45. https://doi.org/ 10.18653/v1 /2020. emnlp-demos.6

[23] Jaradat, S., Nayak, R., Paz, A., & Elhenawy, M. (2024). Ensemble learning with pre-trained transformers for crash severity classification: A deep NLP approach. Algorithms, 17(7), 284.

[24] Saito, T., & Rehmsmeier, M. (2015). The precision–recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLOS ONE, 10(3), e0118432. https://doi.org/10.1371/ journal. pone.0118432