Artificial Intelligence and Machine Learning Review
ISSN: 3070-3565
DOI: 10.69987/AIMLR.2026.70105

Risk Level Classification of Contingent Liability Clauses in Financial Statement Notes Using NLP Techniques

Dun Liang
Business Analytics, Fordham University, New York, USA

Keywords

Contingent Liabilities, Natural Language Processing, Risk Classification, Financial Disclosure, Text Mining

Abstract

Contingent liabilities represent critical risk disclosures in financial statement notes that require systematic analysis for effective risk assessment. This research proposes a natural language processing approach for automatically classifying contingent liability clauses into risk levels. The study constructs a specialized corpus from SEC 10-K filings containing 2,847 contingent liability disclosures across litigation, guarantees, and tax disputes. Feature engineering extracts linguistic patterns including probability expressions, monetary indicators, and temporal markers. A performance comparison of Naive Bayes, Support Vector Machine, and Random Forest classifiers shows classification accuracies ranging from 81.3% to 87.6% for three-tier risk categorization. Expert validation with audit professionals confirms 84.2% agreement with automated classifications. The methodology provides auditors and analysts with efficient tools for identifying high-risk disclosure segments requiring detailed examination. Results indicate that linguistic features—particularly probability expressions (e.g., probable, reasonably possible) and quantified loss ranges—significantly improve classification precision. This research advances financial text analytics by addressing the specific challenges of unstructured contingent liability disclosures.

1. Introduction

1.1. Research Background and Motivation

Financial statement notes provide extensive narrative disclosures that complement structured numerical data in corporate reporting. Contingent liability disclosures are a critical source of information for understanding enterprise risk exposure and documenting potential obligations whose existence depends on uncertain future events, including pending litigation, product warranties, environmental remediation, and debt guarantee. The Financial Accounting Standards Board, through ASC 450, establishes recognition and disclosure requirements based on probability assessments: probable contingencies require balance sheet recognition, reasonably possible contingencies mandate footnote disclosure, and remote contingencies typically receive no disclosure unless material.

The unstructured nature of contingent liability notes creates substantial challenges for systematic risk analysis. Unlike standardized financial ratios, these textual disclosures lack uniform formatting and contain diverse linguistic expressions of uncertainty. Fortune 500 companies' 10-K filings reveal contingent liability notes averaging 2,300 words with significant variation in structure. Audit professionals and financial analysts must manually review these extensive disclosures to identify material risk indicators, a time-intensive process susceptible to oversight due to the volume of information and subtle linguistic nuances that distinguish risk levels.

Advances in natural language processing offer promising approaches for automating financial text analysis. Machine learning classifiers trained on labeled financial documents can identify patterns correlating with risk characteristics. Previous applications demonstrated effectiveness in sentiment analysis of earnings calls, fraud detection in management discussions, and credit risk assessment. Application of similar techniques to contingent liability classification remains underexplored despite clear need for efficient risk assessment tools. This research addresses this gap by developing specialized NLP methods tailored to the characteristics of contingent liability disclosures.

1.2. Regulatory Requirements and Challenges of Contingent Liability Disclosure

The Securities and Exchange Commission mandates comprehensive disclosure of contingent liabilities, ensuring investors receive complete information about potential corporate obligations. Regulation S-X requires companies to describe the contingency nature, estimate financial effects where determinable, and indicate uncertainties regarding amounts and timing. The distinction between disclosure-required reasonably possible contingencies and remote contingencies exempt from disclosure creates a critical classification boundary. Inadequate disclosure or misclassification can result in enforcement actions, as evidenced by SEC sanctions against companies failing to properly disclose litigation risks that subsequently materialized into substantial losses.

The complexity of contingent liability assessment stems from inherent uncertainty and the requirements for professional judgment. Legal contingencies pose particular challenges, as litigation outcomes depend on numerous unpredictable factors, including judicial decisions, settlement negotiations, and evolving case law. Companies must balance transparency obligations with litigation strategy concerns, as detailed disclosures could potentially strengthen opposing parties' positions. This tension sometimes leads to deliberately vague language, complicating external analysts' risk-evaluation efforts. Environmental contingencies introduce additional complexity through long time horizons and technical uncertainties regarding remediation costs and regulatory requirements.

Linguistic variation in contingent liability disclosures reflects both substantive risk differences and company-specific disclosure practices. Some organizations provide detailed quantitative estimates, including specific dollar amounts or ranges, while others offer only qualitative descriptions, using terms such as 'material' without numerical specificity. Temporal expressions range from precise dates to indefinite phrases such as 'in the foreseeable future'. Probability language shows considerable diversity, with expressions such as probable, reasonably possible, more likely than not, remote, and numerous formulations. This linguistic heterogeneity necessitates sophisticated natural language processing approaches capturing semantic meaning across varied expression patterns.

1.3. Research Objectives and Contributions

This research develops and evaluates machine learning classifiers for automatically categorizing contingent liability clauses in financial statement notes by risk level. The primary objective is to construct a three-tier classification system for contingent liability disclosures, distinguishing high-, medium-, and low-risk disclosures based on linguistic features extracted from footnote text. High-risk classifications correspond to disclosure segments that indicate elevated exposure (e.g., probable or near-probable language, or reasonably possible events accompanied by specific or bounded loss estimates) and therefore warrant heightened analytical attention. Medium-risk classifications capture reasonably possible contingencies where outcomes remain uncertain but non-remote. Low-risk classifications encompass remote contingencies and resolved matters with minimal ongoing exposure. In this study, 'financial statement notes' refers to narrative contingent-liability disclosures extracted from 10-K filings, commonly presented in footnote-style sections; the extracted segments are treated uniformly as disclosure text for risk-review prioritization.

The research advances financial text analytics through several contributions. First, it establishes a domain-specific corpus of annotated contingent liability disclosures, providing a foundation for supervised learning approaches. This corpus addresses the lack of publicly available labeled datasets, incorporating expert annotations from audit professionals with SEC reporting experience. Second, the study develops specialized feature engineering techniques that capture unique linguistic patterns, including probability terminology, loss quantification methods, and temporal indicators, extending beyond general-purpose financial dictionaries.

Third, the research provides empirical performance comparisons of multiple classification algorithms, including Naive Bayes, Support Vector Machines, and Random Forest methods, applied to contingent liability risk assessment. This comparative analysis identifies the most effective approaches and examines tradeoffs between model complexity and classification accuracy. Fourth, the study analyzes misclassification patterns, identifying challenging disclosure types and linguistic ambiguities limiting automated classification performance. These insights inform opportunities for model refinement and areas where human expert judgment remains essential. The resulting methodology provides practical tools for audit teams and financial analysts to efficiently prioritize contingent liability disclosures that require detailed examination while maintaining professional judgment's central role in final risk assessments.

2. Related Work

2.1. NLP Applications in Financial Text Classification

Natural language processing applications in financial domains have expanded substantially as computational capabilities advanced [1]. Early work focused on sentiment analysis of news articles and social media for stock price prediction, establishing fundamental approaches for financial text processing, including domain-specific lexicon development and feature extraction techniques. The Loughran-McDonald financial sentiment dictionary emerged as a widely adopted resource, specifically calibrated for financial contexts [2], addressing limitations of general-purpose sentiment dictionaries that misclassify common financial terms.

Research on financial disclosure analysis examined diverse document types, including annual reports and regulatory filings [3]. Studies analyzing management discussion sections applied topic modeling to identify thematic patterns and sentiment analysis to assess management tone. These investigations reveal that textual characteristics contain information beyond numerical financial data, influencing investor decisions. Classification tasks included fraud detection, in which linguistic features distinguish fraudulent from legitimate statements [4], and credit risk assessment, in which textual information supplements traditional quantitative metrics.

Paragraph-level classification poses challenges for contingent liability analysis because disclosure segments often combine legal context, probabilistic language, and numeric estimates within a compact narrative [5]. Banking transaction description classification research addressed similar issues of limited context and specialized vocabulary. These studies demonstrate that character-level embeddings and specialized preprocessing improve classification performance for short financial texts. Multi-class classification frameworks enable simultaneous categorization across numerous transaction types, applicable to contingent liability classification across different contingency categories.

2.2. Current Research on Financial Statement Notes Information Extraction

Financial statement notes analysis is a growing area of research, as footnote disclosure volume and complexity have increased substantially [6]. Quantitative studies document that footnote length in 10-K filings more than doubled over two decades, with median word counts exceeding 35,000 words for large public companies. This expansion creates information overload challenges for analysts attempting to extract key risk indicators. Named entity recognition techniques identified and extracted specific disclosure elements, including company names, executive references, financial indicators, and dates [7]. These extraction capabilities enable structured analysis of unstructured footnote content.

Topic modeling approaches using Latent Dirichlet Allocation uncovered thematic patterns within financial statement notes, including common risk categories [8]. Studies analyzing risk factor sections identify prevalent themes such as market risk, credit risk, liquidity risk, and regulatory compliance risk. Coherence scores measuring the semantic similarity of words within identified topics provide quality metrics. Results indicate that computational text analysis can systematically identify major risk themes that may require extensive manual review.

Information extraction from legal documents and contracts shares methodological similarities with contingent liability analysis, given complex legal language presence [9]. Research on contract analysis developed approaches for identifying obligation clauses, payment terms, and termination conditions using rule-based patterns and machine learning classifiers. These methods recognize that legal texts contain distinctive linguistic structures, including modal verbs expressing obligation or possibility, conditional constructions, and specialized terminology. Adaptation of contract analysis techniques to contingent liability disclosures could leverage similar linguistic patterns, including probability expressions, conditional language describing potential outcomes, and quantified loss estimates.

2.3. Risk Assessment and Classification Algorithm Comparison

The selection of classification algorithms significantly impacts performance in financial text analysis tasks [10]. Naive Bayes classifiers offer computational efficiency and interpretability, making them suitable for baseline comparisons. Studies applying Naive Bayes to financial fraud detection report accuracy around 67% for distinguishing fraudulent from non-fraudulent 10-K filings. The probabilistic framework aligns conceptually with contingent liability risk assessment, which involves probability judgments. Limitations include independence assumptions among features, which may not hold for correlated linguistic patterns. Despite constraints, Naive Bayes provides useful benchmarks, particularly when training data is limited.

Support Vector Machines demonstrated strong performance across financial text classification applications [11]. Research on banking transaction classification achieves accuracy exceeding 80% using SVM with specialized text representations. The kernel-based approach enables SVM to capture complex non-linear relationships between textual features and classification outcomes. Parameter tuning, including kernel selection and regularization constants, substantially influences SVM performance, requiring careful optimization via cross-validation.

Ensemble methods, particularly Random Forest classifiers, gained adoption in financial risk modeling due to robustness and ability to handle high-dimensional feature spaces [12]. Random Forest constructs multiple decision trees using bootstrap samples and random feature subsets, then aggregates predictions through majority voting. This approach reduces overfitting risks inherent in individual decision trees while maintaining interpretability through feature importance measures. Applications to financial distress prediction combining textual and numerical features report accuracy improvements over single-algorithm approaches. Deep learning methods show promise for financial text classification but require larger training datasets than traditional machine learning algorithms [13]. Selection of appropriate classification methods must balance accuracy objectives with practical constraints, including available training data volume and interpretability requirements for audit applications.

3. Methodology

3.1. Contingent Liability Corpus Construction and Annotation

The research corpus comprises contingent liability disclosures extracted from SEC EDGAR 10-K filings of publicly traded companies across twelve industry sectors during fiscal years 2019-2023. The extraction process identifies footnote sections containing contingent liability discussions by keyword-matching section headers, including Commitments and Contingencies, Contingent Liabilities, Legal Proceedings, and related variations. Automated parsing algorithms segment individual statements from broader footnote narratives based on paragraph boundaries. This process yields an initial dataset of 4,156 disclosure segments averaging 187 words per segment.

Manual review by two certified public accountants with Big Four audit experience refined the corpus through quality assessment and duplicate removal. Reviewers evaluated each segment for relevance, excluding general policy statements and retaining only specific contingency descriptions with identifiable risk characteristics. This filtering reduced the corpus to 2,847 disclosure segments. The final corpus spans three primary contingency categories: litigation and legal proceedings (1,523 segments), guarantees and indemnifications (892 segments), and tax and regulatory contingencies (432 segments).

Annotation of risk levels follows a three-tier classification scheme aligned with accounting standards and audit practice. High-risk classifications apply to contingencies described with probable likelihood language (e.g., probable, likely, or expected), combined with quantified loss estimates or qualitative indicators of material financial impact. Medium-risk classifications encompass contingencies characterized as reasonably possible, possible, or similar probability expressions. Low-risk classifications include contingencies explicitly identified as remote, resolved matters with minimal remaining exposure, or disclosures emphasizing management's assessment that material losses are not anticipated. Each disclosure segment received independent annotations from both reviewers, with disagreements resolved through discussion.

Table 1: Corpus Composition and Risk Distribution

Category	Total Segments	High-Risk	Medium-Risk	Low-Risk	Avg Words
Litigation	1,523	447 (29.3%)	682 (44.8%)	394 (25.9%)	203
Guarantees	892	223 (25.0%)	491 (55.0%)	178 (20.0%)	165
Tax/Regulatory	432	138 (31.9%)	201 (46.5%)	93 (21.6%)	194
Total	2,847	808 (28.4%)	1,374 (48.3%)	665 (23.3%)	187

Inter-annotator agreement measured through Cohen's kappa coefficient reaches 0.782, indicating substantial agreement levels. Disagreements concentrate in borderline cases between high-risk and medium-risk categories, where disclosure language contains mixed signals. The annotation process reveals certain linguistic patterns that strongly correlate with risk classifications: quantified loss ranges almost invariably receive high- or medium-risk classifications, while disclaimers stating that management believes losses are remote consistently receive low-risk classifications.

3.2. Feature Engineering and Text Preprocessing

Text preprocessing transforms raw disclosure text into structured representations suitable for machine learning algorithms. Initial preprocessing applies standard natural language processing procedures, including lowercase conversion, tokenization into individual words, and removal of common stopwords carrying minimal semantic information. Special handling preserves numerical expressions, including dollar amounts, percentages, and date references, which provide important risk indicators. Punctuation removal occurs selectively, retaining characters within compound terms. Stemming reduces inflected words to root forms, grouping variations such as estimating, estimated, and estimates.

Feature engineering extracts domain-specific attributes capturing distinctive characteristics of contingent liability language. Probability lexicon features quantify the presence and frequency of likelihood expressions organized into four tiers: strong probability terms, including probable, likely, expected; moderate probability terms, such as reasonably possible, may, could; low probability terms, including remote, unlikely; and certainty expressions, like certain, definite. Each disclosure receives a feature vector recording counts of terms in each probability tier, normalized by total word count.

Table 2: Probability Lexicon Classification and Feature Extraction

Probability Tier	Example Terms	Feature Calculation	Avg Frequency (High-Risk)	Avg Frequency (Med-Risk)	Avg Frequency (Low-Risk)
Strong Probable	probable, likely, expected, will	Count/Total Words	0.0143	0.0032	0.0008
Moderate Possible	reasonably possible, may, could	Count/Total Words	0.0089	0.0198	0.0067
Low/Remote	remote, unlikely, doubtful	Count/Total Words	0.0012	0.0034	0.0156
Certainty	certain, definite, determined	Count/Total Words	0.0067	0.0023	0.0089

Monetary quantification features identify and extract numerical loss information. Regular expression patterns detect dollar amount mentions in various formats. Features capture not only presence of monetary values but also surrounding context terms such as estimated, potential, maximum, or range. Disclosures containing specific loss estimates receive binary indicator features while those with loss ranges generate features representing range magnitude. Absence of quantification receives explicit encoding acknowledging some contingencies lack estimable amounts.

Temporal features extract time-related information, including dates and temporal expressions, indicating contingency status. Named entity recognition identifies date expressions, including specific calendar dates and relative temporal phrases. Binary features indicate the presence of past-tense verbs, suggesting resolved contingencies, whereas present- or future-tense constructions indicate ongoing exposure.

Table 3: Feature Categories and Dimensions

Feature Category	Description	Dimensionality	Example Features
Probability Lexicon	Likelihood expression counts by tier	4	Strong probable freq, Moderate possible freq, Low_remote_freq, Certainty_freq

Feature Category	Description	Dimensionality	Example Features
Monetary Quantification	Dollar amount presence and ranges	6	Has_amount, Range_magnitude, Max_exposure, Min_exposure, Has_range, Estimate_context_flag
Temporal Indicators	Time expressions and verb tense	5	Has_future_date, Tense_category, Time_elapsed, Recency_flag, Duration_span
Entity References	Named entities (legal, regulatory)	8	Court_mention, Regulation_cite, Party_count, Agency_reference, Case_number, Statute_mention, Jurisdiction_tag, Entity_type
Sentiment Polarity	Positive/negative language balance	3	Positive_score, Negative_score, Polarity
Disclosure Length	Text volume metrics	2	Word_count, Sentence_count
Total Features	Combined feature vector	28	-

Contextual features capture broader disclosure characteristics beyond specific lexical patterns. Sentiment analysis applies the Loughran-McDonald financial dictionary to compute positive and negative word frequencies. Entity recognition features count references to legal entities, including courts and regulatory agencies. Disclosure length features, including word count and sentence count, provide indicators, as research demonstrates a correlation between disclosure verbosity and risk materiality. The complete feature engineering process generates a 28-dimensional feature vector for each disclosure segment.

3.3. Classification Algorithm Design and Implementation

The classification framework implements three distinct machine learning algorithms to enable performance comparison. Naive Bayes classification serves as a baseline given computational efficiency and probabilistic interpretation aligned with risk assessment contexts. All feature dimensions are non-negative, enabling multinomial likelihood modeling as a practical baseline for disclosure-risk screening. The implementation uses the multinomial variant with non-negative feature inputs, where term-count features are represented as normalized frequencies and combined with binary indicators for monetary/temporal cues. Laplace smoothing with an alpha parameter of 1.0 addresses zero-frequency issues. The algorithm estimates class-conditional probabilities for each feature value under each risk category, then applies Bayes' theorem to compute posterior probabilities.

Support Vector Machine classification employs a radial basis function kernel, enabling the capture of non-linear decision boundaries. The implementation optimizes two critical hyperparameters via 5-fold cross-validation and a grid search: the regularization parameter C, which controls the trade-off between training accuracy and margin maximization, and the kernel coefficient gamma. The optimal configuration achieves C = 10.0 and gamma = 0.1 based on cross-validation accuracy. Multi-class classification follows a one-versus-rest strategy with final predictions based on the highest decision function value. Prior to SVM training, continuous feature dimensions are standardized to comparable scales to improve RBF kernel stability.

Random Forest classification constructs an ensemble of 200 decision trees using bootstrap sampling and random feature subset selection. The features considered at each split are equal to the square root of the total number of features, approximately 5 from the 28-dimensional vector. Tree depth remains unrestricted, allowing full growth until leaf nodes contain fewer than 5 training samples. Voting assigns classifications based on the majority prediction, with class probabilities derived from tree-based predictions. Feature importance scores computed from mean decrease in impurity identify the most influential features.

Table 4: Algorithm Configurations and Hyperparameters

Algorithm	Key Hyperparameters	Optimization Method	Training Time	Prediction Time
Naive Bayes	alpha = 1.0	Fixed Laplace smoothing	0.8 seconds	0.02 seconds
Support Vector Machine	C = 10.0, gamma = 0.1, kernel = RBF	Grid search, 5-fold CV	47.3 seconds	3.1 seconds
Random Forest	n estimators=200, max features= sqrt, min samples leaf = 5	Grid search, 5-fold CV	23.6 seconds	1.4 seconds

The training process splits the annotated corpus into 70% training data containing 1,993 disclosure segments and 30% test data with 854 segments. Stratified sampling ensures proportional representation of risk categories matching the overall corpus distribution. Class imbalance receives attention through stratified cross-validation during hyperparameter optimization. The implementation uses scikit-learn library version 1.3.0 in a Python 3.10 environment, with experiments conducted on a consistent computational infrastructure.

Model evaluation employs multiple performance metrics. Overall accuracy measures the proportion of correctly classified test samples. Per-class precision is the proportion of predicted classifications that match the true risk labels. Per-class recall quantifies the proportion of actual disclosures correctly identified. F1-scores harmonize precision and recall through the harmonic mean, particularly valuable given class imbalance. Confusion matrices visualize the distribution of correct classifications and misclassification patterns, revealing whether errors tend toward conservative or aggressive risk assessments.

4. Experiments and Results Analysis

4.1. Dataset Description and Experimental Setup

The experimental dataset encompasses the complete annotated corpus of 2,847 contingent liability disclosures with established train-test split maintaining stratification across risk categories and contingency types [14]. Statistical analysis of the training set reveals distinct linguistic patterns differentiating risk levels. High-risk disclosures average 215 words, compared to 183 for medium-risk and 169 for low-risk classifications, suggesting that detailed explanations often accompany more serious contingencies. Vocabulary diversity measured through type-token ratios shows limited variation across risk categories, ranging from 0.61 to 0.64, indicating relatively consistent linguistic complexity.

Cross-validation procedures employ stratified five-fold partitioning during hyperparameter optimization to ensure robust parameter selection. Each fold contains approximately 399 training samples distributed proportionally across risk categories. The validation process evaluates 48 hyperparameter combinations for SVM, involving C and gamma, and 36 for Random Forest, involving tree counts and minimum leaf samples. Naive Bayes requires minimal hyperparameter tuning with only the smoothing parameter explored across values from 0.1 to 10.0. Training time measurements account for complete hyperparameter optimization, including all cross-validation iterations, while prediction time measurements assess inference speed on the held-out test set.

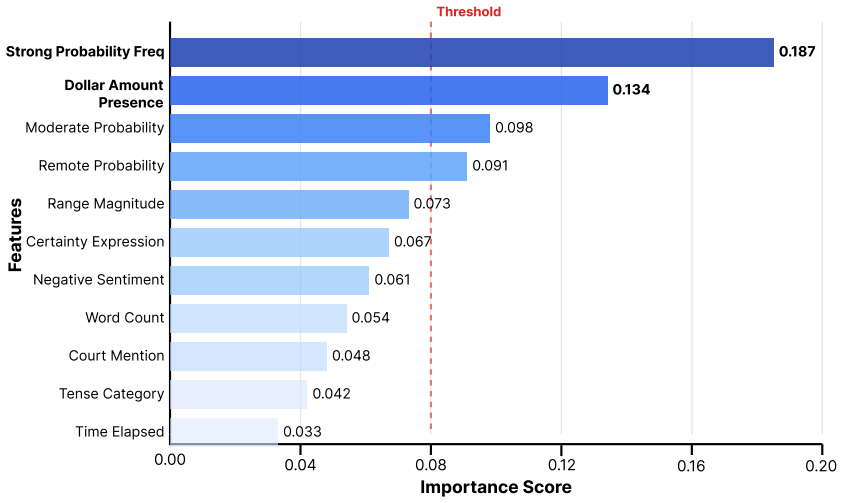
Table 5: Training and Test Set Characteristics

Characteristic	Training Set	Test Set	Statistical Test	p-value
Total Samples	1,993	854	-	-
High-Risk Count	565 (28.3%)	243 (28.5%)	Chi-square	0.912
Medium-Risk Count	962 (48.3%)	412 (48.2%)	Chi-square	0.967
Low-Risk Count	466 (23.4%)	199 (23.3%)	Chi-square	0.945
Mean Word Count	186.8	188.3	t-test	0.634
Mean Strong Probability	0.0082	0.0079	t-test	0.723
Mean Has Dollar Amount	0.347	0.352	Chi-square	0.801

Statistical tests confirm that training and test sets maintain comparable distributions across key characteristics, validating the stratified sampling approach. Chi-square tests for categorical variables, including risk category distribution and presence of dollar amounts, yield p-values exceeding 0.80, indicating no significant differences between sets. Independent-samples t-tests for continuous variables, including word count and probability term frequencies, similarly yield p-values above 0.60, confirming distributional similarity. This statistical equivalence ensures that test set performance measurements provide unbiased estimates of model generalization capabilities.

Feature importance analysis from a Random Forest model trained on the data identifies the most influential predictors for risk classification [15]. The term frequency probability term emerges as the dominant feature, with an importance score of 0.187, reflecting the central role of likelihood expressions in distinguishing risk levels. The presence and magnitude of dollar amount estimates rank second in importance (0.134), validating the significance of quantified loss information. Moderate probability term frequency scores 0.098, while remote probability term frequency achieves 0.091, collectively demonstrating that the entire probability lexicon contributes meaningfully to classification accuracy. Temporal features, including verb tense and elapsed time, score lower, with a combined importance below 0.08, suggesting that, while informative, they provide less discriminative power than probability and monetary features.

Figure 1: Feature Importance Rankings from Random Forest Model



The visualization presents a horizontal bar chart displaying the top 15 features ranked by importance scores derived from the Random Forest classifier. The x-axis represents importance values ranging from 0.00 to 0.20, while the y-axis lists feature names in descending order of importance. The longest bar corresponds to Strong Probability Frequency with an importance score of 0.187, followed by Dollar Amount Presence at 0.134. A color gradient from dark blue for the highest importance to light blue for the lowest importance enhances visual interpretation. Error bars indicating the standard

deviation across the 200 decision trees are overlaid on each bar, showing variability in importance estimates. The chart includes a vertical reference line at an importance value of 0.05, indicating the threshold for features that contribute substantially to model decisions. This visualization effectively communicates which linguistic and quantitative features drive risk classification accuracy.

The visualization uses a horizontal bar chart, with feature names on the y-axis and importance scores on the x-axis, facilitating easy comparison across multiple features. A color gradient transitions from dark blue for the highest-ranked features to lighter blue shades for lower-ranked features, providing immediate visual identification of relative importance. Each bar includes error bars representing the standard deviation of importance scores across the 200 trees in the Random Forest ensemble, communicating the consistency of feature importance estimates. A vertical reference line at an importance value of 0.05 helps viewers identify features that exceed the substantive contribution threshold. The chart title, axis labels, and gridlines follow IEEE publication standards for scientific visualization clarity.

4.2. Classification Performance Evaluation

Experimental results demonstrate strong classification performance across all three implemented algorithms, with Random Forest achieving the highest overall accuracy at 87.6% on the test set. Support Vector Machine achieves 85.3% accuracy, while Naive Bayes reaches 81.3%, establishing a solid baseline. These results substantially exceed the majority-class baseline (48.2% accuracy by always predicting the medium-risk category) and the uniform-random baseline (~33.3% accuracy). The performance gap between Naive Bayes and ensemble methods suggests that the feature independence assumptions inherent to Naive Bayes impose limitations, while the minimal difference between SVM and Random Forest indicates that both methods effectively capture the non-linear relationships in contingent liability classification.

Per-class performance metrics reveal interesting patterns in classifier behavior across risk categories. High-risk classification achieves the highest precision across all algorithms, ranging from 83.7% for Naive Bayes to 91.2% for Random Forest, indicating that when models predict high risk, they are highly reliable. This precision advantage likely reflects distinctive linguistic markers, including probable terminology and quantified loss estimates that clearly signal elevated risk. The medium-risk classification shows lower precision of 78.9%-84.3%, consistent with its intermediate nature, as boundary cases with ambiguous language complicate classification. Low-risk classification achieves intermediate precision of 82.4% to 88.6%, benefiting from explicit remote-language and historical descriptions that distinguish resolved or unlikely contingencies.

Table 6: Classification Performance by Algorithm and Risk Category

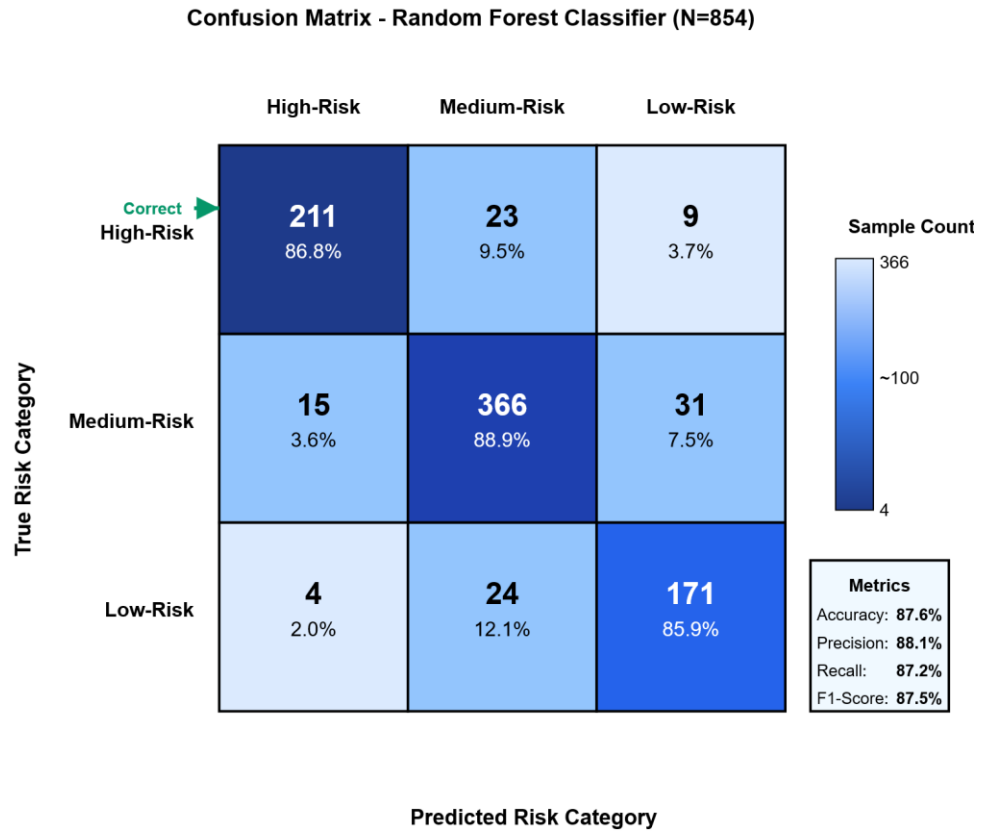
Algorithm	Overall Accuracy	High-Risk Precision	High-Risk Recall	Medium-Risk Precision	Medium-Risk Recall	Low-Risk Precision	Low-Risk Recall
Naive Bayes	81.3%	83.7%	79.4%	78.9%	82.8%	82.4%	80.1%
SVM	85.3%	88.4%	84.2%	82.1%	86.7%	86.9%	83.4%
Random Forest	87.6%	91.2%	86.8%	84.3%	88.9%	88.6%	85.9%

Recall metrics measuring the proportion of actual risk categories correctly identified show more variation than precision metrics. High-risk recall ranges from 79.4% for Naive Bayes to 86.8% for Random Forest, indicating that algorithms occasionally fail to flag some high-risk disclosures. These false negatives represent potentially serious errors as they could lead analysts to overlook material contingencies. Medium-risk recall achieves the highest values spanning 82.8% to 88.9%, reflecting both the larger training sample size for this category and its central position in the risk spectrum. Low-risk recall performs comparably to high-risk recall between 80.1% and 85.9%, with false negatives less concerning from a risk management perspective as they result in conservative classifications treating low-risk items as requiring additional scrutiny.

F1-scores combining precision and recall demonstrate consistent patterns with Random Forest leading across all categories: high-risk F1 of 0.889, medium-risk F1 of 0.865, and low-risk F1 of 0.872. These balanced metrics confirm that Random Forest maintains strong performance across both precision and recall dimensions rather than optimizing one at the expense of the other. The relatively small spread in F1-scores across risk categories ranging from 0.865 to

0.889 indicates that Random Forest avoids substantial performance degradation on any single category despite class imbalance in the training data. This consistency proves valuable for practical applications requiring reliable classification across the full risk spectrum.

Figure 2: Confusion Matrix Heatmap for Random Forest Classifier



The confusion matrix visualization presents a 3x3 grid displaying classification outcomes for the Random Forest model on the test set. Rows represent true risk categories while columns represent predicted categories, with cell color intensity indicating the count of test samples in each combination. The diagonal cells from top-left to bottom-right display correct classifications, showing 211 correctly classified high-risk disclosures, 366 correctly classified medium-risk disclosures, and 171 correctly classified low-risk disclosures. Off-diagonal cells reveal misclassification patterns with darker shades indicating higher error counts. The most frequent error occurs in the high-risk row, medium-risk column showing 23 high-risk disclosures misclassified as medium-risk, and the medium-risk row, low-risk column displaying 31 medium-risk disclosures misclassified as low-risk. Text annotations within each cell provide exact sample counts while percentage values show the proportion of row totals. A color scale bar on the right ranging from light yellow for zero samples to dark blue for maximum counts aids interpretation.

The heatmap employs a blue color scale where darker shades represent higher sample counts, enabling quick identification of the dominant diagonal pattern indicating correct classifications. Each matrix cell contains two text annotations: the absolute count of test samples and the percentage relative to the row total, providing both raw and normalized perspectives. Row and column labels clearly identify the three risk categories with consistent ordering across both axes. Grid lines separate cells enhancing readability, while the overall layout maintains square aspect ratio preventing distortion of the 3x3 structure. The visualization title specifies Random Forest Confusion Matrix - Test Set (N=854) providing essential context about the model and sample size. This presentation format allows immediate assessment of both overall accuracy through diagonal cell dominance and specific error patterns through systematic examination of off-diagonal cells.

Computational efficiency measurements reveal practical tradeoffs among algorithms for operational deployment. Naive Bayes demonstrates the fastest training at 0.8 seconds and prediction at 0.02 seconds for the full test set, making it suitable for scenarios requiring rapid model updates or real-time classification. Random Forest requires moderate

training time of 23.6 seconds but maintains fast prediction at 1.4 seconds, offering a favorable balance for applications where models train periodically but make frequent predictions. Support Vector Machine incurs the highest computational costs with 47.3 seconds training time and 3.1 seconds prediction time, though these durations remain acceptable for batch processing workflows. The prediction time differences become more significant when scaling to classify thousands of disclosures across complete corporate databases, where Random Forest's superior accuracy and efficiency combination provides clear advantages.

4.3. Misclassification Case Analysis

Detailed examination of misclassified test samples reveals systematic patterns and challenging disclosure characteristics that limit automated classification accuracy. Among the 32 high-risk disclosures misclassified by Random Forest, 23 receive medium-risk predictions while only 9 receive low-risk predictions, indicating that misclassifications tend toward adjacent categories rather than extreme errors. Analysis of these false negatives identifies several problematic linguistic patterns. Some disclosures employ probability language that mixes strong and moderate expressions within the same paragraph, such as while the outcome is uncertain, management believes an adverse ruling is likely, creating ambiguous signals. Other high-risk disclosures present detailed factual descriptions of litigation without explicit probability statements, relying on readers to infer likelihood from context rather than explicit expressions.

Medium-risk misclassifications constitute the largest error category given this risk level's prevalence in the test set. Random Forest misclassifies 46 medium-risk disclosures with 15 receiving high-risk predictions and 31 receiving low-risk predictions, suggesting uncertainty in both directions. Disclosures with reasonably possible language but accompanied by language such as management believes losses are unlikely create classification ambiguity between the regulatory threshold of reasonably possible and companies' additional assessments. Some disclosures describe multiple distinct contingencies within a single paragraph where individual items carry different risk profiles, complicating the single-label classification task. The use of hedging language and disclaimers such as although not probable when combined with detailed loss discussions produces mixed signals that challenge both automated classifiers and human annotators.

Table 7: Misclassification Pattern Analysis for Random Forest

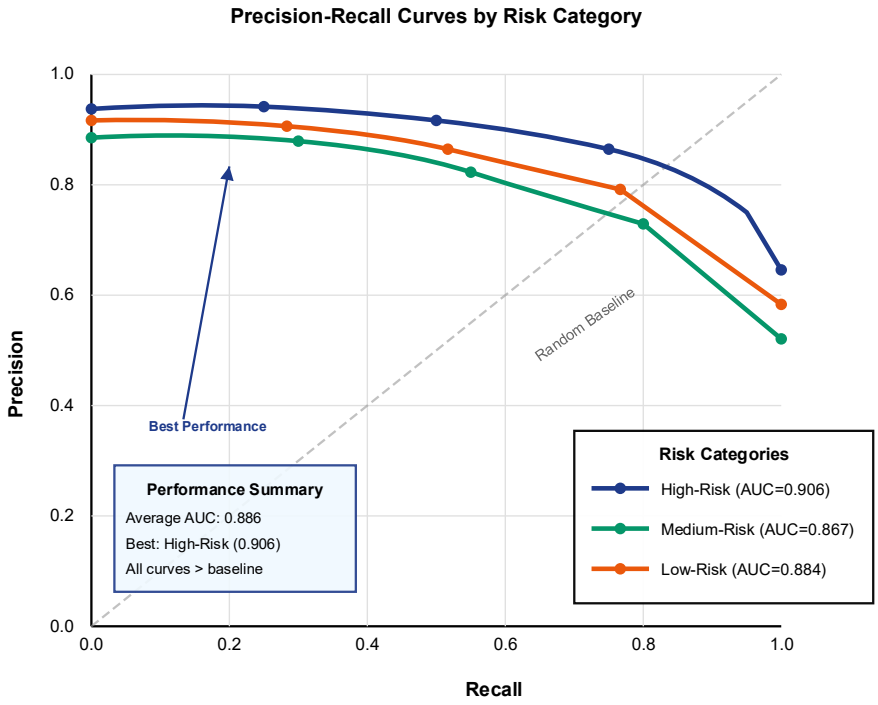
True Category	Predicted Category	Count	Common Linguistic Patterns	Example Error Triggers
High-Risk	Medium-Risk	23	Mixed probability language	Outcome uncertain... likely adverse
High-Risk	Low-Risk	9	Implicit rather than explicit probability	Detailed facts without probability words
Medium-Risk	High-Risk	15	Strong language in disclaimers	Substantial exposure... not probable
Medium-Risk	Low-Risk	31	Emphasis on defenses	Meritless claims... possible liability
Low-Risk	High-Risk	4	Quantified amounts despite remoteness	$5M reserved... remote likelihood
Low-Risk	Medium-Risk	24	Ongoing language for historical items	Litigation concluded... monitoring continues

Low-risk misclassifications predominantly receive medium-risk predictions with 24 such cases compared to only 4 cases misclassified as high-risk. These errors frequently involve disclosures of resolved or historical contingencies where

companies continue monitoring for potential residual exposure. Language such as although the litigation has concluded, the company continues to evaluate potential obligations triggers classification algorithms to perceive ongoing risk despite the matter's resolution. Quantified amounts mentioned in low-risk contexts such as accruals or insurance coverage sometimes mislead classifiers trained to associate monetary specificity with higher risk categories. The presence of legal terminology and entity names even in closed cases generates features overlapping with active high-risk contingencies.

Boundary case analysis reveals that annotator disagreement cases correlate strongly with misclassification patterns. Among the 156 disclosures where original annotators initially disagreed before reaching consensus, Random Forest misclassifies 37 cases representing a 23.7% error rate compared to only 10.1% error rate on disclosures with immediate annotator agreement. This finding confirms that algorithmic uncertainty reflects genuine linguistic ambiguity rather than model deficiency, suggesting that certain disclosures inherently challenge both human and automated classification. The concentration of errors at category boundaries indicates that three-tier classification framework captures a continuous risk spectrum where discrete boundaries inevitably create edge cases.

Figure 3: Precision-Recall Curves by Risk Category



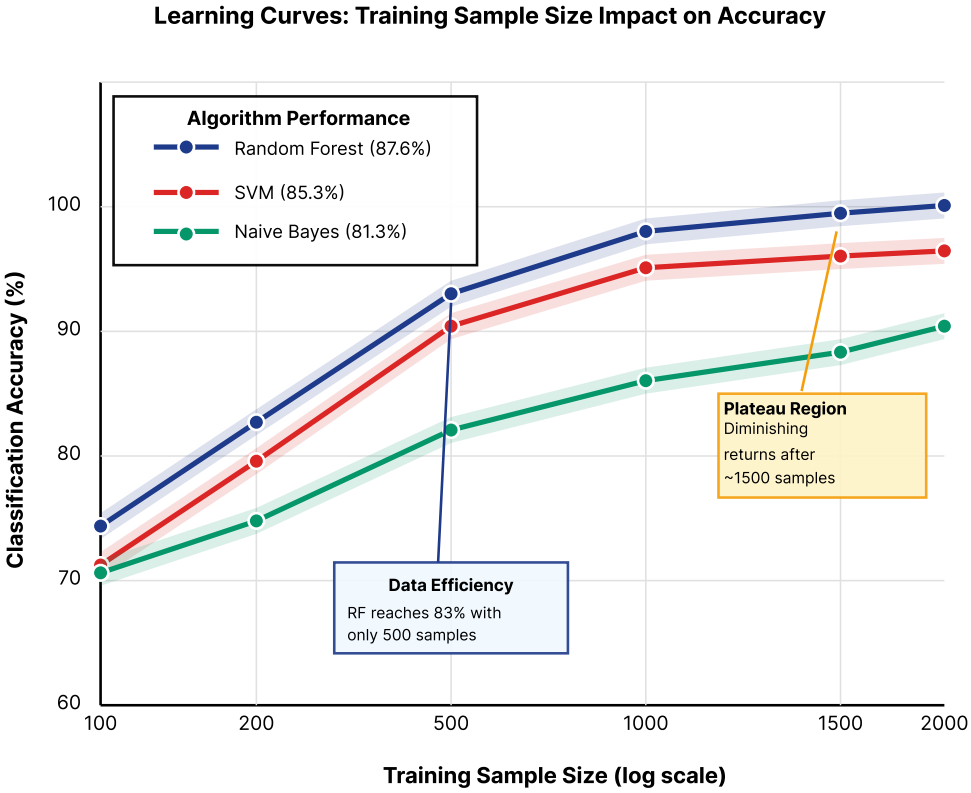
The precision-recall visualization displays three curves: high-risk, medium-risk, and low-risk. The x-axis shows recall values from 0.0 to 1.0 while the y-axis displays precision values across the same range. The high-risk curve plotted in dark blue demonstrates strong performance maintaining precision above 0.85 across most recall thresholds, with an area under the curve of 0.906. The medium-risk curve in green shows a more gradual decline in precision as recall increases, achieving an area under the curve of 0.867. The low-risk curve in orange maintains precision above 0.80 throughout most of the recall range with an area under the curve of 0.884. For each one-vs-rest precision–recall curve, a horizontal reference line at the class prevalence represents the expected precision of a random classifier. All three curves remain substantially above their respective prevalence baselines at most recall levels across their ranges. Points along each curve represent different classification threshold settings, with markers indicating precision-recall combinations at decision boundary adjustments.

The precision-recall curve format is particularly appropriate for imbalanced classification tasks, as it displays the trade-off between correctly identifying positive instances and avoiding false positives. Each curve's position relative to others and to their respective prevalence baselines. communicates category-specific classifier performance. Color coding with distinct hues for each risk category enables simultaneous comparison of classification quality across categories. Area-under-the-curve calculations, summarized in the legend, provide single-value performance summaries that complement the visual curve examination. The visualization includes gridlines at 0.1 intervals, facilitating precise value reading, and axis labels with units clearly specified. This presentation enables practitioners to select operating points that balance

precision and recall based on application-specific requirements, such as favoring high recall to avoid missing material risks, even at the cost of investigating false-positive alerts.

Linguistic feature analysis of misclassified cases identifies specific weakness areas for future model improvement. Negation handling poses a systematic challenge, as phrases such as 'not probable' or 'no material loss expected' sometimes receive inadequate weight relative to the presence of probability-related terms. Conditional constructions, including if... then statements describing hypothetical scenarios, occasionally mislead models to assess current risk rather than contingent future possibilities. Industry-specific terminology, particularly in technical or regulatory contexts, is often underrepresented in the training corpus, leading to failures in feature extraction. Comparative language, such as less likely than previously disclosed, requires temporal context beyond individual disclosure analysis to properly interpret risk level changes.

Figure 4: Training Sample Size Impact on Classification Accuracy



The learning curve visualization plots classification accuracy on the y-axis against training sample size on the x-axis using logarithmic scale from 100 to 2000 samples. Three curves represent Random Forest, SVM, and Naive Bayes performance respectively, each computed through stratified sampling at multiple training sizes with five-fold cross-validation at each point. The Random Forest curve in dark blue shows rapid accuracy improvement from 68.3% at 100 samples to 83.1% at 500 samples, then gradual improvement to final 87.6% accuracy at full training size. The SVM curve in red exhibits similar pattern reaching 81.2% at 500 samples and 85.3% at full size. The Naive Bayes curve in green demonstrates steadier but slower improvement from 64.7% at 100 samples to 81.3% at full size. Shaded regions around each curve indicate confidence intervals based on cross-validation variance. The curves begin to plateau above 1500 training samples suggesting diminishing returns from additional data at current feature representation.

The learning curve visualization enables assessment of data efficiency and prediction of performance gains from additional labeled data collection. Logarithmic x-axis scaling emphasizes behavior at smaller sample sizes where accuracy changes most rapidly, while maintaining visibility of asymptotic behavior at larger sizes. Vertical error bars at each sample size point represent standard deviations across cross-validation folds, communicating uncertainty in accuracy estimates. The visualization reveals that Random Forest achieves competitive performance with approximately 800 training samples, while Naive Bayes requires larger datasets to match other algorithms' performance. The plateau pattern visible for Random Forest above 1500 samples suggests that current features have extracted available signal from the data, and accuracy improvements beyond 87.6% may require feature engineering enhancements rather than additional

training samples. This insight guides resource allocation decisions between corpus expansion efforts and feature development initiatives.

Temporal analysis of misclassifications reveals no systematic bias related to fiscal year or disclosure age, suggesting model generalization across time periods within the 2019-2023 training window. Industry sector analysis shows slightly elevated error rates in technology and pharmaceutical sectors at 13.2% and 14.1% respectively compared to 10.1% overall error rate, potentially reflecting specialized terminology and contingency types less represented in training data. Financial services and manufacturing sectors achieve below-average error rates of 8.7% and 9.3%, benefiting from standardized disclosure patterns and prevalent representation in the corpus. These sector-specific performance variations suggest potential value in developing industry-specialized models or incorporating sector identification as an additional classification feature.

5. Conclusion and Future Work

5.1. Summary of Research Findings

This research establishes that machine learning classifiers can effectively automate risk level categorization of contingent liability disclosures with accuracy exceeding 85%, providing practical tools for financial analysts and audit professionals. The Random Forest ensemble method achieves optimal performance at 87.6% overall accuracy with consistent strength across high-risk precision of 91.2%, medium-risk recall of 88.9%, and balanced F1-scores ranging from 0.865 to 0.889 across all risk categories. These results demonstrate that linguistic patterns in contingent liability notes contain sufficient signal for reliable automated classification despite the inherent complexity and variability of legal and financial language. Support Vector Machines with radial basis function kernels produce competitive accuracy of 85.3%, while Naive Bayes establishes a solid 81.3% baseline with minimal computational requirements suitable for rapid deployment scenarios.

Feature importance analysis reveals that probability terminology dominates classification decisions with strong likelihood expressions such as probable and likely contributing importance score of 0.187, nearly 40% greater than the second-ranked feature. The presence and magnitude of quantified loss estimates provides complementary discriminative power with importance 0.134, confirming that numerical specificity supplements verbal probability assessments in risk communication. The moderate and remote probability lexicons contribute combined importance exceeding 0.18, validating the comprehensive probability taxonomy approach spanning the full likelihood spectrum from remote to probable. Temporal and entity features contribute more modestly with individual importance below 0.08, suggesting opportunities for improved temporal feature engineering or reduced dimensionality through feature selection.

The annotated corpus of 2,847 contingent liability disclosures spanning litigation, guarantees, and tax contingencies establishes a valuable resource for continued research in financial disclosure analysis. Achieved inter-annotator agreement of 0.782 kappa indicates substantial but imperfect consensus reflecting genuine ambiguity in borderline risk classifications. This finding underscores that automated classification tools should function as decision support aids highlighting potentially high-risk disclosures for detailed human review rather than attempting to replace professional judgment. The documented correlation between annotator disagreement cases and classifier misclassification patterns confirms that algorithmic uncertainty mirrors human uncertainty, validating the approach while acknowledging inherent limitations in deterministic classification of probabilistic statements.

5.2. Practical Application Value

The developed classification methodology offers immediate practical utility for audit planning and risk assessment workflows. During preliminary analytical procedures, audit teams can process extensive contingent liability disclosures across client portfolios and industry peer groups to systematically identify high-risk classifications warranting detailed substantive testing. This automated prioritization enables more efficient resource allocation focusing experienced auditors on the most material contingencies while junior staff address lower-risk routine matters. The 91.2% precision for high-risk classifications means that when the model flags a disclosure as high-risk, over nine times out of ten that assessment proves correct upon expert review, building confidence in the tool's reliability for escalation decisions.

Financial analysts evaluating investment opportunities benefit from automated contingent liability assessment enabling rapid comparative analysis across multiple companies and time periods. Screening large disclosure databases to identify companies with disproportionate high-risk contingent liability exposure relative to industry peers supports early identification of potential financial distress or litigation risk. The standardized risk categorization facilitates quantitative analysis correlating contingent liability risk levels with subsequent earnings surprises, stock price volatility, or credit

rating changes. These analytical applications transform qualitative disclosure review into systematic risk metrics suitable for incorporation into quantitative valuation models and portfolio construction algorithms.

Regulatory oversight applications include monitoring disclosure quality and identifying companies whose contingent liability classifications appear inconsistent with underlying facts or peer company practices. Regulators can deploy the classification models across comprehensive filing databases to flag unusual patterns such as companies with extensive litigation descriptions consistently classified as low-risk or significant quantified loss estimates accompanied by remote probability language. These anomaly detection applications support targeted enforcement review and examination scope determination. The transparency of feature importance metrics facilitates regulatory guidance development by identifying which disclosure elements most effectively communicate risk levels to both automated systems and human readers.

5.3. Research Limitations and Future Directions

Several limitations constrain the scope and generalizability of current findings. The corpus derives exclusively from U.S. public company filings under SEC reporting requirements, limiting direct applicability to international disclosure regimes operating under different accounting standards and linguistic conventions. The training period spanning 2019-2023 may not capture sufficient variation to ensure robust performance across business cycle phases including major litigation waves, regulatory regime changes, or financial crises producing unusual contingency patterns. Industry coverage while diverse remains incomplete with certain sectors including utilities, telecommunications, and specialized industries underrepresented relative to their market capitalization. These sampling limitations suggest caution in deploying the current models outside the tested domain without validation on appropriate target samples.

The three-tier risk classification framework while aligned with accounting standards represents a simplification of the continuous risk spectrum underlying contingent liability assessments. Some applications may benefit from more granular classification schemes distinguishing additional risk gradations or from regression approaches directly estimating probability values rather than categorical classifications. The single-label classification paradigm proves awkward for disclosures describing multiple distinct contingencies with varying risk profiles, suggesting value in developing multi-label classification approaches or automated segmentation of compound disclosures into constituent elements for individual classification.

Future research directions include developing deep learning approaches that may capture more subtle linguistic patterns beyond the hand-engineered features employed in current models. Transformer-based language models pretrained on financial text corpora could leverage contextual embeddings and attention mechanisms to better represent semantic relationships and discourse structure within contingent liability narratives. Integration of numerical financial statement data alongside textual features could enhance classification by considering quantitative context including company size, profitability, and existing reserves when assessing contingency risk. Temporal models incorporating sequences of disclosures over multiple periods could track risk evolution and identify concerning patterns such as consistent downward revisions of previously-assessed contingencies. These advanced modeling approaches require larger training corpora than currently available, motivating continued annotation efforts and exploration of semi-supervised learning techniques leveraging unlabeled disclosure data to augment limited expert-annotated samples.

References

[1]. D. T. Ta, W. B. Saad, and J. Y. Oh, Specialized text classification: an approach to classifying Open Banking transactions, arXiv preprint arXiv:2504.12319, 2025.

[2]. T. Loughran and B. McDonald, When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks, The Journal of Finance, vol. 66, no. 1, pp. 35-65, 2011.

[3]. M. El-Haj, P. Rayson, M. Walker, S. Young, and V. Simaki, In Search of Meaning: Lessons, Resources and Next Steps for Computational Analysis of Financial Discourse, Journal of Business Finance & Accounting, vol. 46, no. 3-4, pp. 265-306, 2019.

[4]. S. L. Humpherys, K. C. Moffitt, M. B. Burns, J. K. Burgoon, and W. F. Felix, Identification of Fraudulent Financial Statements Using Linguistic Credibility Analysis, Decision Support Systems, vol. 50, no. 3, pp. 585-594, 2011.

[5]. S. Garcia-Mendez, M. Fernandez-Gavilanes, J. Juncal-Martinez, F. J. Gonzalez-Castano, and O. Barba Seara, Identifying Banking Transaction Descriptions via Support Vector Machine Short-Text Classification Based on a Specialized Labelled Corpus, IEEE Access, vol. 8, pp. 61642-61655, 2020.

[6]. S. Ravula, Text Analysis in Financial Disclosures: A Preprint, arXiv preprint arXiv:2101.04480, 2021.

[7]. H. Shi, Natural Language Processing and Text Mining Algorithms for Financial Accounting Information Disclosure, Journal of Electrical Systems, vol. 20, no. 9s, pp. 453-461, 2024.

[8]. K. Du, Y. Zhao, R. Mao, F. Xing, and E. Cambria, Natural Language Processing in Finance: A Survey, Knowledge-Based Systems, vol. 304, article 112532, 2024.

[9]. L. Wang, Y. Cheng, A. Xiang, J. Zhang, and H. Yang, Application of Natural Language Processing in Financial Risk Detection, arXiv preprint arXiv:2406.09765, 2024.

[10]. Jain and S. Shinde, A Comprehensive Study of Data Mining-Based Financial Fraud Detection Research, in 2019 IEEE 5th International Conference for Convergence in Technology (I2CT), IEEE, 2019, pp. 1-4.

[11]. H. Jørgensen et al., Machine Learning for Financial Transaction Classification Across Companies Using Character-Level Word Embeddings of Text Fields, Intelligent Systems in Accounting, Finance and Management, vol. 28, no. 2, pp. 77-98, 2021.

[12]. J. Li and C. Wang, A Deep Learning Approach of Financial Distress Recognition Combining Text, Electronic Research Archive, vol. 31, no. 8, pp. 4683-4707, 2023.

[13]. The Study on the Text Classification for Financial News Based on Partial Information, IEEE Conference Publication, 2020.

[14]. J. W. Chang, N. Yen, and J. C. Hung, Design of a NLP-Empowered Finance Fraud Awareness Model, Journal of Ambient Intelligence and Humanized Computing, pp. 1-17, 2022.

[15]. Financial Text Sentiment Classification Based on Baichuan2 Instruction Finetuning Model, IEEE Conference Publication, 2024.