iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** IV **Month of publication:** April 2026

DOI: https://doi.org/10.22214/ijraset.2026.80641

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue IV Apr 2026- Available at www.ijraset.com

Automated Term Sheet Validation: Methods, Datasets, and Challenges

Ishant Deoghare[1], Bhajneet Singh Hooda[2], Prakash Deshmukh[3], Abhishek Dukare[4], Vaishnavi Chavhan[5]
Department of Information Technology Pune Institute of Computer Technology Pune, India

Abstract: ***The validation of term sheets in venture capital, private equity, and syndicated lending transactions is a crit-ical task that ensures accuracy, completeness, and regulatory compliance. Traditional manual validation processes are time-consuming, error-prone, and difficult to scale. Recent advances in artificial intelligence, natural language processing (NLP), and large language models (LLMs) have enabled automated approaches to term sheet validation. This paper provides a comprehensive review of methods and datasets for automated term sheet validation, emphasizing the transition from rule-based systems and classical machine learning to transformer-based architectures and hybrid models. We examine challenges including dataset limitations, semantic understanding, cross-domain adaptation, and ethical considerations. Furthermore, we present future research directions aimed at developing robust, in-terpretable, and scalable validation systems for complex financial and legal documents.***

Index Terms: Term Sheet Validation, Legal Tech, Natural Language Processing, Contract Analysis, Clause Detection, Com-pliance Automation

I. INTRODUCTION

Term sheets are foundational documents in investment and lending transactions, outlining key terms, conditions, and expectations between parties before definitive agreements are executed. In venture capital (VC), private equity (PE), and syndicated finance, term sheets specify valuation, liquidation preferences, anti-dilution provisions, board composition, and investor rights [1], [2].

Their accurate validation is essential for risk assessment, due diligence, and regulatory compliance. Manual term sheet validation involves legal experts review-ing documents clause-by-clause to ensure completeness, con-sistency, and alignment with industry standards and regulatory requirements. This process is labor-intensive, prone to human error, and challenging to scale as transaction volumes increase [3], [4]. The complexity of legal language, jurisdiction-specific variations, and evolving regulatory frameworks compound these difficulties.

The emergence of artificial intelligence (AI) and NLP tech-nologies offers promising solutions to automate and enhance term sheet validation. Early approaches relied on rule-based systems and keyword matching, while recent advancements leverage deep learning, transformer-based models (e.g., BERT, GPT), and hybrid architectures that combine textual analysis with graph-based representations [5]–[7]. These methods can identify and classify clauses, detect inconsistencies, and flag missing or non-compliant provisions with increasing accuracy. This paper surveys the landscape of automated term sheet validation, reviewing methodologies from classical machine learning to state-of-the-art LLMs. We examine publicly avail-able datasets, evaluation metrics, and benchmarking results. Additionally, we identify persistent challenges including data scarcity, semantic ambiguity, domain adaptation, and ethical considerations. By synthesizing existing research, we aim to provide a roadmap for developing more effective, transparent, and scalable term sheet validation systems.

II. LITERATURE SURVEY

A. Systematic Reviews and Surveys

Several comprehensive reviews have aggregated existing methods for legal document analysis and contract validation. Industry guidance from the Loan Market Association (LMA) and European Leveraged Finance Association (ELFA) high-lights investor concerns regarding term sheet completeness, noting frequent omissions of material terms such as financial definitions, restricted payment baskets, and ESG disclosures [3]. Legal tech surveys emphasize the growing adoption of AI for contract review, clause extraction, and compliance checking [8], [9].

©IJRASET: All Rights are Reserved

B. Classical Machine Learning Methods

Early automated contract analysis employed traditional ma-chine learning techniques. Support Vector Machines (SVMs) with TF-IDF features achieved robust performance for text classification tasks [10], [11]. Decision trees, random forests, and Na¨ıve Bayes classifiers were applied to detect risky clauses and categorize contract types [12]. While effective for structured datasets, these methods struggle with semantic nuances, context-dependent interpretation, and complex clause interdependencies.

C. Deep Learning and Transformer-Based Models

The advent of deep learning revolutionized contract anal-ysis. Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) improved feature extraction from legal text [13]. Transformer-based models, particularly BERT and its variants, achieved state-of-the-art results in clause clas-sification and contract understanding [5]. The Contract Under-standing Atticus Dataset (CUAD) provided expert-annotated contracts across 41 clause categories, serving as a benchmark for model evaluation [5]. Studies show BERT-based models achieve F1-scores ranging from 75% to 96% depending on clause complexity [14].

D. Clause Extraction and Classification

Recent research focuses on automatic clause extraction and classification. ConReader explores implicit relations in contracts for clause extraction [6]. ClauseRec proposes a recommendation framework for AI-aided contract authoring [7]. ContractNLI introduces a dataset for document-level natural language inference in contracts [26]. Benchmarking studies reveal proprietary models (e.g., GPT-4) outperform open-source alternatives in correctness (F1: 0.641) and output effectiveness [14]. Enterprise solutions like Sirion and Icertis report clause extraction accuracy exceeding 94% [15].

E. Legal NLP and Contract Analysis Tools

Specialized NLP tools for legal contracts have emerged. LexCheck and VerifAI leverage generative AI for contract review, achieving 73-87% accuracy in clause identification [16], [17]. Studies on construction contracts demonstrate NLP-based risk assessment models achieve 83% accuracy while re-ducing review time by 76% [18]. Financial NLP applications, including BloombergGPT, showcase domain-specific language models trained on extensive financial corpora [19].

F. Venture Capital and Investment Agreement Analysis

Research on VC term sheets emphasizes negotiation dy-namics, risk allocation, and value effects of specific provisions [20], [21]. Automated investment agreement analysis using AI tools has gained traction, with applications in due diligence, portfolio optimization, and decision support [22], [23]. Studies highlight the importance of explainability and transparency in AI-driven investment analysis [24].

III. METHODOLOGY

A. Data Collection and Annotation

Automated term sheet validation requires high-quality an-notated datasets. We propose collecting term sheets from public repositories (e.g., SEC EDGAR), anonymizing sensitive information, and engaging legal experts to annotate clauses by category (e.g., valuation, liquidation preference, anti-dilution, board composition, investor rights, protective provisions).

B. Preprocessing and Feature Engineering

Preprocessing steps include tokenization, stop word re-moval, and normalization. Advanced feature engineering em-ploys word embeddings (Word2Vec, GloVe) and contextu-alized representations (BERT, RoBERTa). Domain-specific embeddings trained on legal and financial corpora enhance semantic understanding [19], [25].

C. Model Architecture

We propose a hybrid architecture combining:

- Transformer-based encoders (BERT, Legal-BERT) for clause representation
- Classification heads for multi-label clause categorization
- Graph neural networks (GNNs) to model clause inter-dependencies
- Rule-based validators for compliance checking against industry standards

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

D. Training and Validation

Models are trained using supervised learning with cross-entropy loss. We employ 10-fold cross-validation to ensure ro-bustness. Hyperparameter tuning uses grid search or Bayesian optimization. Transfer learning from pre-trained legal language models accelerates convergence and improves generalization [5], [27].

E. Evaluation Metrics

Performance is assessed using:

- Precision, Recall, F1-Score: Measure clause detection accuracy
- Jaccard Similarity: Evaluate output effectiveness
- False Negative Rate: Assess missed clauses (critical in legal contexts)
- Processing Time: Quantify efficiency gains

IV. DATASETS AND BENCHMARKS

A. CUAD (Contract Understanding Atticus Dataset)

CUAD contains 510 commercial legal contracts with expert annotations for 41 clause types. It serves as a primary bench-mark for contract analysis tasks [5]. Recent studies report F1-scores of 0.53-0.64 across models [14].

B. FinQA and Financial Datasets

FinQA provides question-answering pairs over finan-cial reports, supporting numerical reasoning tasks [28]. BloombergGPT leverages a 363 billion token financial dataset [19].

C. Term Sheet Collections

Industry sources provide term sheet templates and best practice guides:

- LMA/ELFA Best Practice Guide for Term Sheet Com-pleteness [3]
- NVCA Model Legal Documents [29]
- Startup India and LetsVenture Term Sheet Templates [30]

D. Evaluation Benchmarks

ContractEval benchmarks 19 LLMs on contract review tasks, evaluating correctness, output effectiveness, and false negative rates [14]. PROVBENCH assesses legal provision recommendation and conflict detection [31].

V. EMPIRICAL RESULTS

Table I summarizes empirical results from recent term sheet and contract validation studies.

TABLE I
SUMMARY OF TERM SHEET AND CONTRACT VALIDATION STUDIES

Sr.	Study/System	Method	Year	Accuracy/F1	Key Findings
1	CUAD Benchmark	BERT-based models	2021	F1: 0.75-0.96	Expert-annotated dataset with 41 clause types; foundation for con-tract AI
2	ContractEval	19 LLMs (GPT-4, Qwen, LLaMA)	2025	F1: 0.53-0.64	Proprietary models outperform open-source; size has diminishing returns
3	Sirion CLM	Hybrid small+large LMs	2025	94.2%	Enterprise-grade accuracy with ex-plainability and security

4	LegalGraph AI	Custom AI engine	2024	87%	75% speed improvement; 10% ac- curacy gain over baseline
5	Concord AI	Trained on 10k+ con- tracts	2025	98%	92 min to 26 sec per contract; production-verified results
6	Construction Contract NLP	NLP+ML risk assess- ment	2025	83%	76% time reduction; effective for complex technical contracts
7	ConReader	Implicit relation ex- traction	2022	SOTA on CUAD	Captures interdependenci es between clauses
8	ClauseRec	AI-aided authoring	2021	High precision	Framework for clause recommen- dation in drafting
9	ContractNLI	Document-level NLI	2021	Baseline set	Inference over full contract context
10	BloombergGPT	50B param LLM on finance data	2023	SOTA on finance tasks	Largest domain-specific dataset (363B tokens)

VI. CHALLENGES

A. *Dataset Limitations*

Most datasets are English-centric, limiting applicability to multilingual and cross-jurisdictional scenarios. Annotated term sheet datasets are scarce, requiring costly expert involvement. Dataset imbalance (e.g., rare clauses) affects model perfor-mance [14].

B. *Semantic and Contextual Understanding*

Legal language is nuanced, with context-dependent mean-ings. Models struggle with ambiguity, satire, implicit state-ments, and jurisdiction-specific interpretations. Even advanced LLMs exhibit ”hallucinations” or confident incorrect outputs [32].

C. *Scalability and Computational Complexity*

Transformer models require substantial computational re-sources. Real-time validation at scale demands efficient ar-chitectures. Model compression, knowledge distillation, and lightweight variants are needed [14].

D. *Explainability and Trust*

Legal practitioners require transparent, interpretable sys-tems. Black-box models hinder adoption. Techniques like SHAP, attention visualization, and rule-based hybrid systems enhance explainability [10], [15].

E. *Cross-Domain and Cross-Jurisdictional Adaptation*

Models trained on US contracts may underperform on European or Asian jurisdictions. Transfer learning and domain adaptation strategies are essential [3].

F. *Ethical and Governance Concerns*

Automated systems risk amplifying biases, stifling authentic negotiation, and enabling censorship. Fairness, accountability, and stakeholder collaboration are critical [33].

VII. FUTURE DIRECTIONS

1) *Multilingual and Cross-Cultural Models:* Develop multilingual corpora and cross-lingual transfer learning to support diverse legal systems and languages.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

2) *Efficient Architectures:* Lightweight transformers, model compression, and edge deployment enable real-time validation.
3) *Explainable and Auditable Systems:* Integrate interpretability frameworks (SHAP, LIME) and rule-based validators for transparent decision-making.
4) *Low-Resource Domain Adaptation:* Few-shot learning, synthetic data generation, and transfer learning address data scarcity in specialized domains.
5) *Integration with Workflow Systems:* Seamless integration with CLM platforms, CRM systems, and deal management tools enhances practical adoption
6) *Adversarial Robustness:* Develop defenses against adversarial attacks and manipu-lated inputs to ensure system reliability.
7) *Collaborative Governance:* Engage researchers, practitioners, regulators, and civil so-ciety to establish ethical guidelines and accountability frame-works.

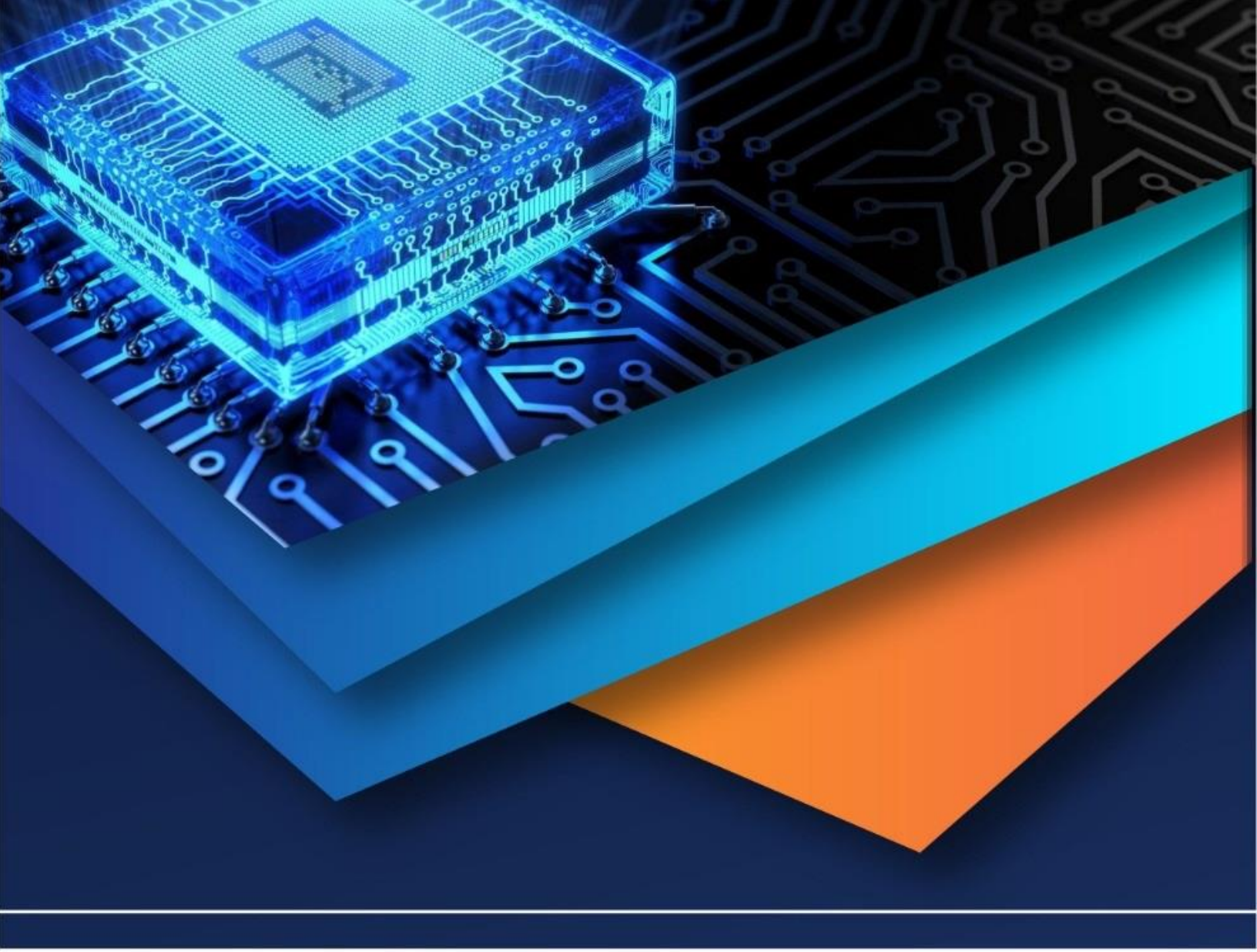
VIII. CONCLUSION

Automated term sheet validation represents a transforma-tive opportunity to enhance accuracy, efficiency, and com-pliance in financial and legal transactions. This paper has surveyed methods from classical machine learning to state-of-the-art transformer-based models, reviewed key datasets and benchmarks, and identified persistent challenges. While significant progress has been made, challenges including data scarcity, semantic ambiguity, scalability, and explainability re-main. Future research must prioritize multilingual capabilities, lightweight architectures, interpretable systems, and ethical governance. By addressing these challenges, the legal tech community can develop robust, trustworthy, and scalable vali-dation systems that empower legal professionals and enhance transaction integrity.

REFERENCES

[1] Wall Street Prep, "Venture Capital Term Sheet (VC) — Format + PDF Template," 2024.
[2] HSBC Innovation Banking, "Venture Capital Term Sheets: The Basics," 2025.
[3] Loan Market Association and European Leveraged Finance Association, "Best Practice Guide for Term Sheet Completeness," 2021.
[4] I. Dikmen et al., "Automated construction contract analysis for risk and responsibility assessment using NLP and ML," *Information Processing & Management*, 2025.
[5] Atticus Project, "CUAD: An Expert-Annotated NLP Dataset for Legal Contract Review," 2021.
[6] Y. Zhou et al., "ConReader: Exploring Implicit Relations in Contracts for Contract Clause Extraction," *EMNLP*, 2022.
[7] "ClauseRec: A Clause Recommendation Framework for AI-aided Con-tract Authoring," *EMNLP*, 2021.
[8] LegitQuest, "The Future of Banking Compliance Lies in Legal Tech," 2025.
[9] Mitratech, "Transforming How Legal Teams Stay Ahead: AI for Com-pliance," 2025.
[10] C. J. Mahoney et al., "A Framework for Explainable Text Classification in Legal Document Review," 2019.
[11] S. Gupta, "Analysis of Automatic Text Classification of Legal Docu-ments," 2022.
[12] M. H. Kazemi, "Application of NLP-based models in automated detec-tion," 2025.
[13] D. Song, "Multi-label legal document classification: A deep learning approach," 2022.
[14] S. Liu et al., "ContractEval: Benchmarking LLMs for Clause-Level Legal Risk Identification," 2025.
[15] Sirion Labs, "2025 Clause Extraction Accuracy: Sirion vs Open-Source LLMs," 2025.
[16] LegalGraph AI, "Legal Contract Review: Overcoming the Accuracy Hurdle in Legal AI Space," 2024.
[17] SpotDraft, "7 Best Practices to Use NLP for Contracts," 2023.
[18] "An Empirical Study on Contract Review and Risk Assessment," *IEEE*, 2025.
[19] "BloombergGPT: A Large Language Model for Finance," 2023.
[20] "Trust and Control: The Value Effect of Venture Capital Term Sheet Pro-visions as Risk Allocation Tools," *Michigan Business & Entrepreneur-ship Law Review*, 2020.
[21] "Term sheet trends in the venture capital market," *Nature Biotechnology*, 2003.
[22] Agreements.AI, "Investment Agreements with AI," 2024.
[23] LeewayHertz, "AI for Investment Analysis: Use Cases, Benefits and Approaches," 2024.
[24] Harvard Business Review, "AI in Professional Services," 2024.
[25] John Snow Labs, "How Is NLP Used In Finance? The Role of Natural Language Processing," 2025.
[26] X. Wang et al., "ContractNLI: A Dataset for Document-level Natural Language Inference for Contracts," 2021.
[27] J. Devlin et al., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," 2019.
[28] "FinQA: A Dataset of Numerical Reasoning over Financial Data," *EMNLP*, 2021.
[29] NVCA, "Model Legal Documents," 2025.
[30] Startup India, "Term Sheet (Equity) Template," 2023.
[31] "PROVBENCH: A Benchmark of Legal Provision Recommendation and Conflict Detection," *ACL*, 2025.
[32] Stanford University, "Report on Hallucination in Large Language Mod-els," 2024.
[33] D. Mouratidis et al., "From Misinformation to Insight: Machine Learning Strategies for Fake News Detection," 2025.

©IJRASET: All Rights are Reserved



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)