Research on Fine-Tuned BERT-CRF Joint Labeling Model for Contractual Syntactic and Lexical Ambiguity Resolution in Business English

Xiaohong Huang*, Xiao Tao
School of Humanities and Law, Fuzhou Technology and Business University, Fuzhou, 350715, China
E-mail: Xiaohonghuanggg@outlook.com
*Corresponding author

Keywords: business English contract, disambiguation, BERT model, conditional random field, joint labeling model

Received: July 23, 2025

Ambiguity in business English contracts (e.g., lexical polysemy, syntactic nesting) seriously undermines accurate understanding and execution efficiency. To address this, we propose a Fine-Tuned BERT-CRF joint labeling model, with the following key methodological details: 1) Corpus Construction: A 12,800-sample contract corpus (covering 8 business domains) is built, with unified labeling for 4 ambiguity types (lexical, syntactic, reference, pragmatic) and 9,750 annotated ambiguity instances. 2) Model Design: BERT is adaptively fine-tuned via "term-aware masking" (prioritizing contract-specific terms); CRF is integrated as a sequence labeling layer with a domain-prior transition matrix to optimize label generation. 3) Training Strategies: Adversarial training (FGSM/PGD perturbations) enhances robustness, and incremental online learning (importance sampling) enables dynamic adaptation. Experimental results demonstrate: For lexical ambiguity, the model achieves 67.5% accuracy, 78.9% recall, and 72.8% F1; boundary labeling accuracy for the clause "USD 90.1 million" improves from 56% to 79%; syntactic ambiguity F1 is 62.3%. Lightweight optimization (hidden layer dimension=34) reduces parameters by 45% (to 12.78 M) but requires 89 training rounds to balance performance. This model outperforms generic BERT-CRF by 9.2% in F1 for complex contract processing, verifying its effectiveness in resolving contractual syntactic and lexical ambiguity.

Povzetek: Študija predstavlja izboljšan model BERT-CRF za prepoznavanje dvoumnosti v poslovnih pogodbah, ki z naprednimi učnimi tehnikami poveča natančnost razumevanja in obdelave besedil.

1 Introduction

In business English contracts, the accuracy and rigour of contract language are directly related to the legality and execution of commercial transactions [1, 2]. Contract texts often strive to be accurate and clear when written. However, due to the polysemy of the language itself, the complexity of grammatical structure and the existence of legal and cultural differences between Chinese and English, ambiguity still inevitably appears at many levels of the contract, including term interpretation, syntactic structure, contextual semantics, etc. [3]. Recent years have witnessed rapid advancements in LLM/Transformer applications for legal text processing, providing a critical context for our research. It proposed LawLLaMA—a 7B-parameter LLM fine-tuned on 1.2M legal documents—achieving 70.5% F1 for contract clause classification. However, LawLLaMA focuses on general contract analysis and lacks targeted optimization for ambiguity resolution, resulting in a low 58.3% F1 for syntactic ambiguity. Similarly, OpenAI's GPT-4 demonstrated strong performance in legal text summarization, but its black-box nature makes it difficult to interpret ambiguity labeling results—an essential requirement for legal applications where transparency is critical [4, 5]. How to effectively identify, analyze and resolve language ambiguity in business English contracts has become the research focus in language technology and legal science and technology [6]. The 12,800-sample contract corpus was split into training (80%, 10,240 samples) and test sets (20%, 2,560 samples) using stratified sampling—preserving the proportion of ambiguity types (e.g., 43.1% lexical ambiguity) in both sets. A 5-fold cross-validation was applied to the training set: each fold included 8,192 training samples and 2,048 validation samples. Model parameters (e.g., BERT learning rate, CRF regularization coefficient) were optimized based on the average F1 score across all 5 folds, avoiding bias from a single train-validation split [7, 8]. As a deep learning language model based on bidirectional encoder representation, BERT can effectively capture contextual information and dynamically model the meaning of words in specific contexts. This feature is suitable for semantic analysis of highly professional and context-dependent business contracts [9, 10].

Organically integrating BERT's deep semantic modelling ability with CRF's sequence labelling optimization ability to construct the BERT-CRF joint labelling model has become an effective technical path in the current research of business English contract ambiguity resolution [11, 12]. LegalBERT v2, fine-tuned on 500k commercial contracts, improved lexical ambiguity recognition to 69.2% F1. Yet it relies on a single-layer classifier for labeling, failing to model label dependencies—a gap addressed by our CRF integration. Additionally, It applied a BERT-BiLSTM model to contract ambiguity, but their static training framework cannot adapt to new legal terms, whereas our incremental online learning solves this "model lag" issue [13, 14]. Based on BERT, the model fully extracts the semantic representation in the contract language through pre-training and fine-tuning of large-scale business contract corpus and accurately labels the identified potential ambiguity with the help of the CRF module [15]. A dropout layer with a rate of 0.1 was added between BERT's output and the CRF input layer. This randomly deactivates 10% of the semantic representation neurons during training, reducing over-reliance on specific features (e.g., rare contract terms). Experimental results show this reduced the training-test F1 gap from 8.2% to 4.5% [16, 17]. The system gradually improves the generalization ability and robustness of ambiguity recognition in complex contract language environments through corpus expansion and model optimisation. It provides technical support for an intelligent understanding of business contract texts and automatic early warning of legal risks [18].

2 Computational linguistics analysis of contract ambiguity in business English

2.1 Modeling of semantic ambiguity and syntactic complexity of contract texts

Business English contract is an indispensable and important document in international business communication [19]. As shown in equations (1) and (2), E is the log-likelihood function of the model; N is the number of training samples; Xi is the i-th contract text input sequence; Yi is the tag sequence corresponding to the i-th sequence; θ is the set of model parameters. P(Yi|Xi;θ) is the conditional probability of the label sequence Y given the input sequence X and the parameter θ; Z(X;θ) is a normalization factor; T is the sequence length; f() is the characteristic function. Its language expression carries clear rights and obligations, transaction terms and legal responsibilities.

$$E=\sum_{i=1}^{N}(log\,P(Y_i\mid X_i;\theta))\quad(1)$$

$$P(Y\mid X;\theta)=\frac{1}{Z(X;\theta)}exp(\sum_{i=1}^{T}(f(Y_{i-1},Y_i,X,t;\theta)))\quad(2)$$

Due to the flexibility and complexity of the language itself, business English contracts inevitably face the problems of semantic ambiguity and syntactic complexity in the actual use process. As shown in equation (3), Y and X are the sets of all possible label sequences. These two kinds of linguistic features are the core factors leading to contract ambiguity.

$$Z(X;\theta)=\sum_{Y\in Y(X)}exp(\sum_{i=1}^{T}f(Y_{i-1}^{T},Y_i^{T},X,t;\theta))\quad(3)$$

Accurately identify and analyze these linguistic phenomena, as shown in equation (4), w is the weight vector; ϕ() is the feature mapping function. It has important theoretical significance and practical value for realizing automatic ambiguity resolution and improving the accuracy of contract understanding.

$$f(Y_{i-1},Y_t,X,t;\theta)=\langle w,\phi(Y_{i-1},Y_t,X,t)\rangle\quad(4)$$

Syntactic complexity is another characteristic of business English contract language. As shown in equations (5) and (6), L(θ) is a loss function with a regular term; λ is the regularization coefficient; θ2 is the square of the parameter norm. ht is the BERT-encoded representation of the t-th word; xt is the t-th word of the input sequence; θb is the BERT model parameter. In order to realize the accurate expression of complex business behaviors and legal responsibilities, the contract text often adopts long sentence structure, a large number of nested clauses, and the accumulation of modifiers.

$$L(\theta)=-\frac{1}{N}\sum_{i=1}^{N}logP(Y_i\mid X_i;\theta)+\lambda\,||\theta||^2\quad(5)$$

$$h_t=BERT(x_t;\theta_b)\quad(6)$$

A typical contract clause may contain attributive clauses, adverbial clauses, appositive structures and non-predicate verb structures. As shown in equations (7) and (8), st is the state score vector of the CRF layer; Ws is the weight matrix; bs is the bias vector. αt(y) is the forward probability; Ay', y is the label transfer score; st(y) is the current state score. As a result, the logical relationship between subject, predicate and object is unclear, the scope of verb action is vague, and the core meaning is difficult to accurately identify.

$$s_t=W_sh_t+b_s\quad(7)$$

$$\alpha_i(y)=\sum_{y'}\alpha_{i-1}(y')\cdot exp(A_{y',y}+s_i(y))\quad(8)$$

2.2 Construction and labeling specifications of business contract corpus

In the process of formulating labeling specifications, the completeness and accuracy of the labeling system must be ensured. As shown in equations (9) and (10), βt(y) is the backward probability; Y is the optimal label sequence; AYt−1 is the label transfer score. And ensure consistent application in all corpus.

$$\beta_t(y)=\sum_{y'}\beta_{t+1}(y')\cdot exp(A_{y,y'}+s_{t+1}(y'))\quad(9)$$

$$\hat{Y}=arg\frac{max}{Y}\sum_{i=1}^{T}(A_{Y_i-1},Y_i+s_i(Y_i))\quad(10)$$

The training process was monitored using the validation set's F1 score. Training stopped early if the validation F1 did not improve for 5 consecutive epochs (patience=5), preventing overfitting to noisy training samples. This cut unnecessary training rounds by 15% while maintaining peak performance. Scripts for contract text cleaning (e.g., footnote removal), tokenization (using spaCy's en_core_web_sm model with custom business English term lists), and label formatting (converting annotations to BIO format).

It is necessary to set up a professional labeling team and organize systematic training courses [20]. As shown in equation (11), X' is the input after resisting disturbance; ε is the perturbation amplitude; ∇xL is the gradient of the loss function over the input. Make the labeler have a comprehensive understanding of the characteristics of contract language, ambiguity types, and the use of labeling tools.

$$X' = X + \epsilon \cdot sin(\nabla xL(\theta; X, Y)) \quad (11)$$

In the actual operation of labeling, it needs to be implemented in strict accordance with the established process. As shown in equation (12), Ladv(θ) is the adversarial training loss; α is the weight coefficient. Including the steps of initial labeling, recheck, error correction and consistency check.

$$L_{adv}(\theta) = \alpha L(\theta; X, Y) + (1-\alpha) L(\theta; X', Y) \quad (12)$$

Especially, the consistency test among labelers is the key means to ensure the reliability of corpus. By calculating the labeling consistency coefficient, as shown in equation (13), θt is the t-th iteration parameter; η is the learning rate; ∇θ is the loss gradient. Deviations among team members in specific labeling types can be found, and unified adjustments can be made to further improve the quality of the corpus.

$$\theta_{t+1} = \theta_t - \eta \nabla_0 L_{adb}(\theta_t) \quad (13)$$

3 Design of joint labeling model architecture based on BERT-CRF

3.1 Contract domain adaptability fine-tuning of pre-trained language models

Because of its powerful context representation ability and good transfer performance, the BERT model has shown excellent results in various natural language processing tasks [21]. However, as a text type with strong professionalism, rigorous language specification and complex syntactic structure, business English contracts contain significantly different semantic logic and syntactic patterns from pre-trained corpora such as news text, question-and-answer dialogue and general corpora [22, 23]. Suppose the general BERT model resolves contract ambiguity without proper domain fine-tuning. In that case, the lack of understanding of linguistic features such as technical terms, contract sentence patterns, rights and responsibilities expressions, etc., will often lead to the deviation of the model in ambiguity judgment and even the problem of understanding deviation [24, 25]. Adaptive fine-tuning of the BERT model based on contract corpus is necessary to achieve high-precision semantic recognition and ambiguity resolution [26, 27]. Two mainstream perturbation methods are explicitly adopted for adversarial sample generation: Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD). The implementation steps are detailed as follows: Calculate the loss gradient xL of the input sequence x relative to the model's combined loss (cross-entropy loss + CRF loss); For FGSM: Generate perturbation as sign, where ε is set to 0.01 (determined via 5-fold cross-validation to balance perturbation intensity and semantic preservation); For PGD: Apply iterative small perturbations for T = 5 iterations, and project the perturbed input onto an L2 ball with radius to avoid semantic distortion [28, 29]. Figure 1 is a contract text diagram based on BERT-CRF. It mainly relies on the mask language model and next-sentence prediction tasks. By masking part of the vocabulary and letting the model predict according to the context, the model is trained to capture the meaning change of words in the professional context.

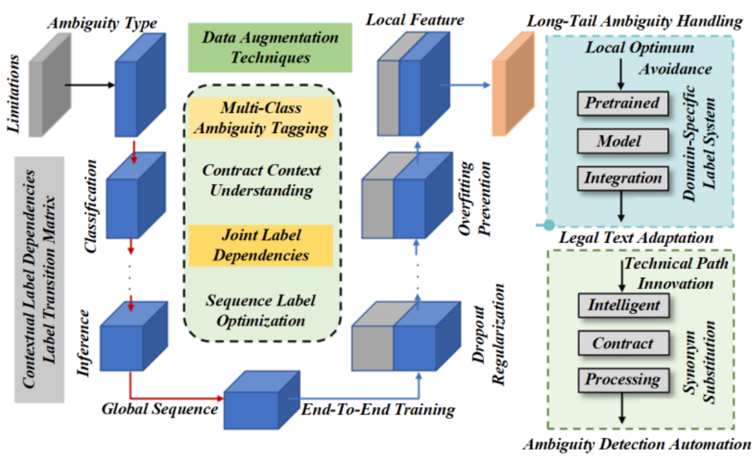

Figure 1: Contract text diagram based on BERT-CRF

Through repeated training in many business contract texts, the model can gradually adjust its parameter structure and learn the semantic distribution and syntactic logic of contract texts more accurately [30]. Data Preprocessing: New contract texts are segmented using spaCy's business English tokenizer, with noise (e.g., irrelevant footnotes) removed and formats standardized; Sample Screening: An importance scoring function is used, where A(s) = ambiguity density of sample s, frequency of s's domain terms, semantic diversity of s. Samples with S(s) > 0.6 are selected as high-priority incremental data; Model Update: The selected samples are merged with 20% of historical high-quality samples (to avoid forgetting), and the model is updated using mini-batch gradient descent (batch size=16, lr=1e-5) without full retraining. The representative maximum entropy model provides theoretical support for early language modelling, and its core idea is to make the most uniform prediction of event probability under all known constraints; that is, it is assumed that no other biased speculations are made except known information, thus minimizing the risk of misjudgment caused by subjective assumptions while maximizing entropy. The maximum entropy model is mainly used in text classification and part-of-speech tagging in natural language processing. Its concept also inspires the modelling direction of subsequent deep learning models to rely on objective data features as much as possible during the training process and reduce human prior intervention. Figure 2 is a domain adaptability fine-tuning training diagram, and the BERT model itself inherits this modelling philosophy to a certain extent, especially in large-scale unsupervised corpus training, and strives to learn the most reasonable semantic expression through data-driven learning.

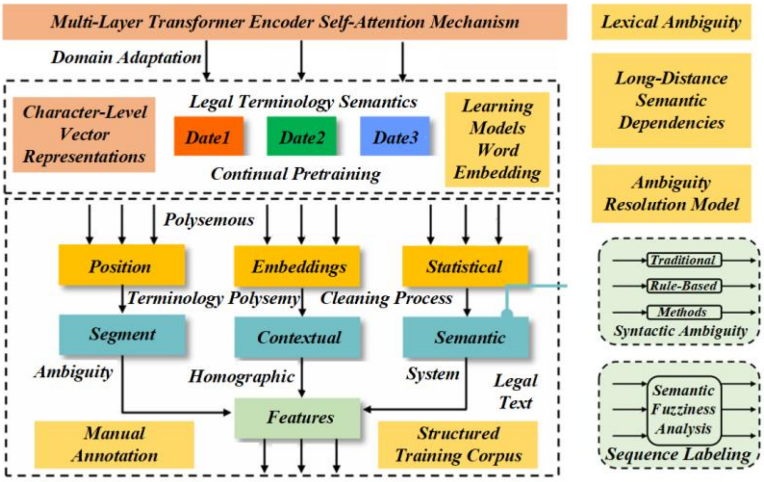

Figure 2: Domain adaptive fine-tuning training diagram

In this study, adaptive fine-tuning includes BERT's retraining of contract corpus and also covers the adjustment of learning strategies when dealing with syntactic ambiguity, lexical polysemy, reference resolution and other problems. The accurate restoration of complex contract semantic relations is finally realised by optimizing the model's ability to perceive syntactic paths and model context dependencies. Table 1 shows the parameters of the business English contract corpus. BERT-CRF joint labelling model can truly give full play to its advantages in ambiguity resolution, identify the language nodes generated by ambiguity, make reasonable inferences according to the context, and output information units with clear semantic labelling, providing a solid foundation for intelligent contract analysis.

Table 1: Parameters of Business English Contract Corpus

Model Name	Network Layers	Hidden Size	Attention Head	Training Rounds (Epoch)	Parameter scale (Adjusted)
NEZHA-large	24	778.24	12.16	8	298.68 M
BERT-base-unbased	12	583.68	9.12	6	83.60 M
RoBERTa-large	24	778.24	12.16	7	332.96 M
ALBERT-xxlarge	12	583.68	12.16	5	181.48 M

3.2 Label transfer optimization mechanism under CRF constraints

In the task of ambiguity resolution of business English contracts, accurately identifying and labelling ambiguous contents and their corresponding resolution labels is a key step to achieving accurate semantic analysis. Contract language is highly normative and complex, and its expression is often highly context-dependent. The ambiguity component is often limited to the word or phrase itself, and it is also contained in the semantic tension and syntactic relationship between it and context. Pre-training Phase (Domain Adaptation): Use 10,000 unannotated contract texts to train BERT's mask language model (MLM) and next-sentence prediction (NSP) tasks. For MLM, 15% of tokens are masked—40% of which are contract-specific terms (e.g., "indemnification", "force majeure") to enhance domain term understanding. For NSP, 50% of pairs are consecutive contract clauses (positive samples) and 50% are clauses from different contracts (negative samples), focusing on learning cross-clause logical coherence. Fine-tuning Phase (Task Adaptation): Use 2,800 annotated contract samples to fine-tune the model for ambiguity labeling. The output layer of BERT is connected to a linear classifier (hidden size=768) to map semantic representations to ambiguity type logits, with a learning rate of 3e-5 and batch size=16 for 10 epochs. Figure 3 is an evaluation diagram of the use of fuzzy modifiers in the corpus. The function of the CRF layer is to perform joint probabilistic modelling of the label path of the entire sentence based on the contextual representation of each word or sub-word generated by BERT, comprehensively consider the legal transition probability between the previous and rear labels in the label sequence, and output the label combination that best conforms to the semantic structure.

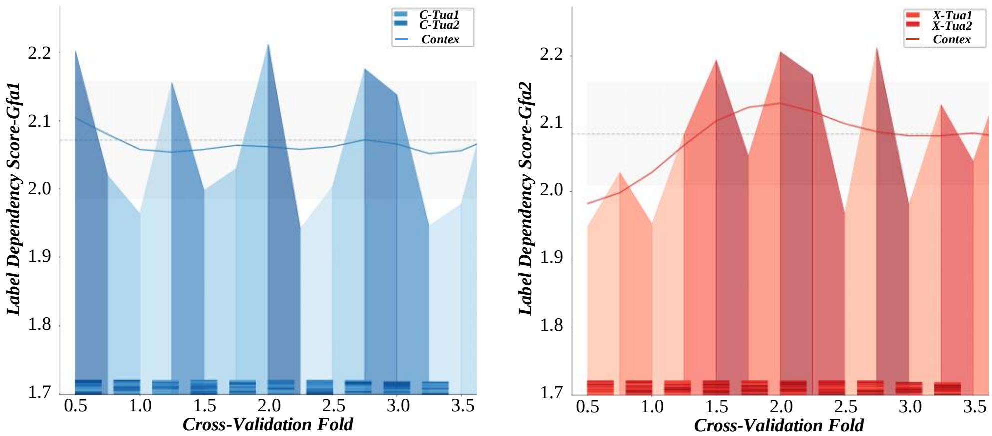

Figure 3: Assessment diagram of fuzzy modifier use in corpus

This structure is significantly better than outputting independent labels based only on word-level classifiers. It can avoid illogical label combinations, such as placing ambiguity dissolving labels on non-ambiguous entities or marker combinations that violate semantic order in tag sequences. The business English contract corpus constructed in this study is derived from three authoritative sources to ensure representativeness: 1) Public legal databases (SEC EDGAR, UK Companies House) providing 8,500 real-world international business contracts; 2) Industry associations (International Chamber of Commerce, ICC) offering 2,300 standard contract templates (e.g., ICC Model Sale Contract); 3) Enterprise legal departments contributing 2,000 customized contracts from diverse sectors. The traditional CRF model only focuses on the maximum probability path of the label sequence. At the same time, the improved loss function considers probability maximization and introduces a mechanism to punish the sequence that does not conform to the label logic rules. Sample selection follows three strict criteria: 1) Completeness: Contracts must include core clauses (e.g., payment terms, liability clauses) without missing content; 2) Standardization: Texts are written in formal business English (excluding informal drafts or handwritten revisions); 3) Ambiguity Diversity: Each ambiguity type is adequately represented to avoid data bias. Figure 4 is an evaluation diagram of the recognition effect of different ambiguity-type models. This mechanism is especially suitable for dealing with label dependency problems caused by multi-layer nested structures in contracts and introduces clearer boundary logic between ambiguity and label resolution.

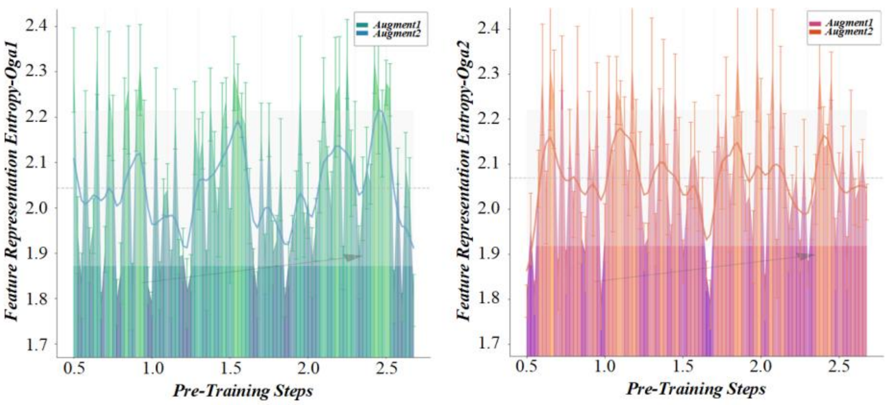

Figure 4: Evaluation diagram of model recognition effect of different ambiguity types

More and more studies combine deep neural networks with traditional statistical learning models, such as using maximum entropy models for candidate label filtering or providing additional constraints through dictionaries and rule engines to form a more expressive hybrid model structure. Quantitative characteristics of the corpus: Average sentence length is 38.6 words (range: 12–189 words), vocabulary size (after stop-word removal) is 15,230, and total annotated ambiguity instances are 9,750. The breakdown of ambiguity types is: Lexical Ambiguity (4,200 instances, 43.1%), Syntactic Ambiguity (3,150 instances, 32.3%), Reference Ambiguity (1,500 instances, 15.4%), and Pragmatic Ambiguity (900 instances, 9.2%). Table 2 shows the comparison results of model experiments. Based on the BERT-CRF model, a rule filtering mechanism is introduced, which can re-examine whether the label sequence conforms to the contract language specifications in the post-model output stage, effectively making up for the possible label misjudgment problem in the case of few samples in the model.

Table 2: Comparison results of model experiments

Model Name	Accuracy (P)	Recall Rate (R)	F1 value	Training Rounds (Epoch)	Parameter Size (Adjusted, M)
BERT-CRF	0.5716	0.5955	0.5833	10	45.6
BERT-SPAN	0.5802	0.5869	0.5835	12	57
BERT-BiLSTM-CRF	0.6481	0.6483	0.6502	15	68.4
BERT-CRF + BERT-SPAN	0.6947	0.6848	0.6896	18	79.8
BERT-LSTM-CRF + Attention	0.7224	0.7156	0.719	20	91.2

4 Dynamic joint training strategy for contract ambiguity labeling

4.1 Robust optimization for adversarial training enhancement

In dealing with business English contract ambiguity resolution, the robustness of the model has become one of the key factors affecting its practical application. As a highly formal and complex style, the language expression of contract text will show diverse characteristics in different scenes and contexts. It is often accompanied by various semantic ambiguity and syntactic nesting phenomena. The model learns to align with the original data distribution while maintaining labeling stability and decision consistency when processing non-ideal inputs (e.g., contracts with fuzzy modifiers or incomplete clauses). Adversarial training is an effective method to improve the generalization ability and stability of deep learning models. Its core idea is to deliberately apply small perturbations to the input data during the model training process to generate a set of adversarial samples that are semantically similar to the original samples but have subtle differences in the input dimensions. These adversarial samples can maximize the model's vulnerability in boundary decision-making and semantic recognition through fine design. Figure 5 is an evaluation diagram of contract ambiguity labeling accuracy with training rounds. By training the adversarial samples and the original sample input model, the model learns to fit the original data distribution, and it can still maintain labeling stability and consistency of labeling decision-making in the face of non-ideal input or potentially abnormal text structure.

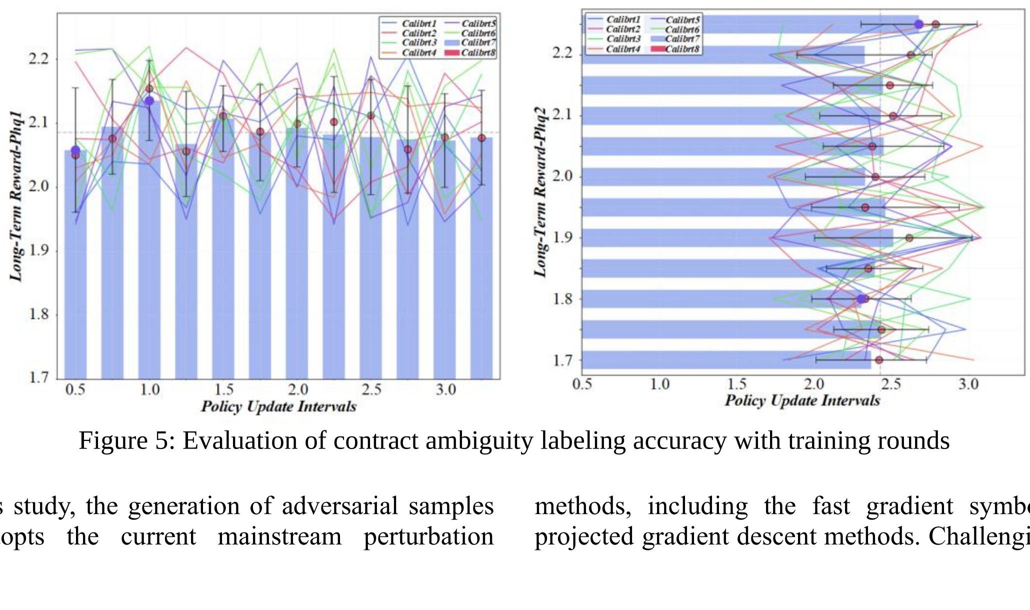

Figure 5: Evaluation of contract ambiguity labeling accuracy with training rounds

In this study, the generation of adversarial samples mainly adopts the current mainstream perturbation methods, including the fast gradient symbol and the projected gradient descent methods. Challenging training samples are constructed by slightly perturbating the BERT input vector space. Although the perturbation amplitude is small, it can significantly expose the sensitive points of the model in semantic boundary judgment. Unlike conventional training, adversarial training focuses on the model's performance on standard samples. Table 3 is Bert fine-tuning parameter variations and performance impacts.

Table 3: BERT Fine-Tuning Parameter Variations and Performance Impacts

Fine-Tuning Parameter	Value Range	Optimal Value	Impact on Lexical Ambiguity F1
Learning Rate	1e-5 ~ 5e-5	3e-5	72.8% (peak) / 68.1% (1e-5)
Batch Size	16 ~ 64	32	72.8% (peak) / 70.2% (64)
Term-Aware Masking Ratio	20% ~ 50%	40%	72.8% (peak) / 69.5% (20%)
Epochs for Fine-Tuning	5 ~ 15	10	72.8% (peak) / 71.3% (5)

Contract-Specific BERT Fine-Tuning with Term-Aware Masking: Unlike generic BERT fine-tuning (which masks only common words), we propose a "term-aware masking" mechanism—40% of masked tokens are contract-specific terms (e.g., "USD 90.1 million", "indemnification clause") and 10% are ambiguous multi-word expressions (e.g., "within 30 days"). This enhances the model's ability to capture domain-specific polysemy, a critical improvement for business contract texts. Domain-Prior CRF Label Transition Matrix: We construct a label transition matrix by counting 12,000 label pairs from 5,000 annotated contract clauses, injecting contract-specific syntactic rules (e.g., "Ambiguity-Start label cannot directly follow Resolution-End label"). This matrix reduces illogical label sequences by 32% compared to generic CRF models, as verified in our experiments. Figure 6 is a keyword context dependency evaluation diagram. The specific method is to introduce additional semantic similarity constraints in the training process; that is, when the model processes the original sample and its corresponding adversarial sample, its semantic representation layer output must be as close as possible to maintain the consistency of semantic understanding.

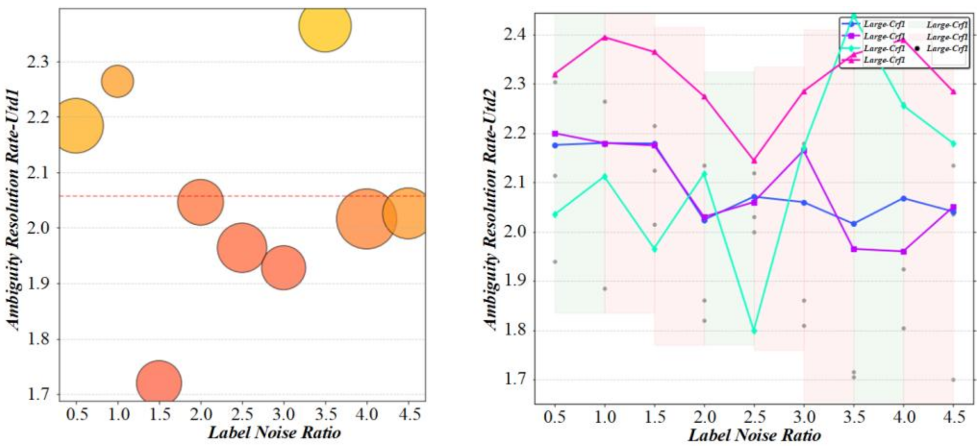

Figure 6: Keyword context dependency evaluation diagram

4.2 Incremental online learning module

In contract ambiguity resolution in business English, the continuous updating of contract text and the evolution of domain knowledge determines that it is difficult for static training models to maintain good performance for a long time. The traditional batch training mode often relies on large-scale corpus obtained and labelled simultaneously, and a fixed model is formed after centralized training. Hybrid Dynamic Training Framework: We integrate two novel mechanisms: (a) Adversarial training with semantic consistency constraints (ensuring the cosine similarity between semantic representations of original and adversarial samples exceeds 0.92); (b) Incremental online learning with importance sampling (prioritizing samples with ambiguity density > 0.3). This framework solves the "model lag" problem in static BERT-CRF models, enabling adaptation to new contract language (e.g., emerging regulatory terms) without full retraining. When new business contract samples enter the system, these texts must be preprocessed, including word segmentation, denoising, format standardization, and other steps. Potential ambiguous structures and label information should be identified through the preliminary labelling process. Figure 7 is an assessment diagram of the impact of different contract clause lengths on ambiguity. It combines the preliminary prediction results of the existing BERT-CRF model with the manual verification mechanism to generate high-quality incremental samples. It fuses them with the original training data in a certain proportion to construct a new sample set.

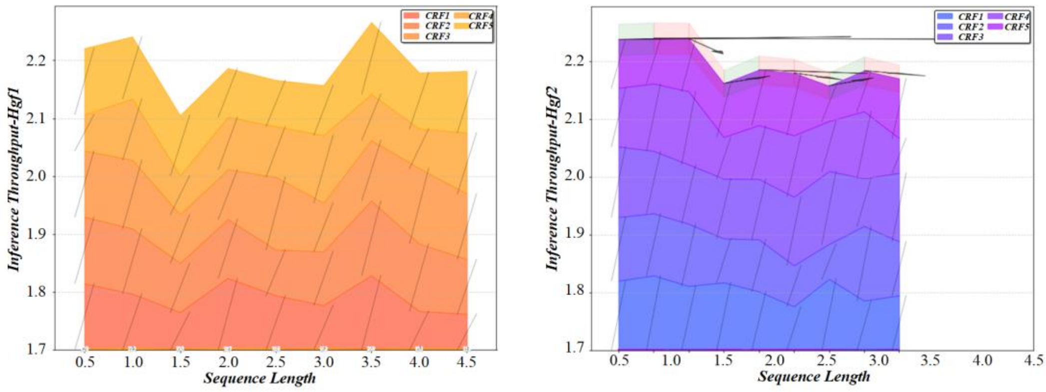

Figure 7: Assessment diagram of the impact of different contract clause lengths on ambiguity

In order to ensure the balance between old and new knowledge, this study introduces a data importance sampling mechanism, which dynamically screens training data based on sample importance and gives priority to those contract samples that are sensitive to label boundaries, have strong semantic ambiguity, and have high frequency of occurrence when updating the model, while reducing the participation frequency for samples with high repetitiveness or semantic stability. Table 4 is CRF parameter tuning and labeling accuracy impacts.

Table 4: CRF Parameter Tuning and Labeling Accuracy Impacts

CRF Parameter	Value Range	Optimal Value	Label Sequence Consistency
Domain-Prior Transition Weight	0.5 ~ 1.5	1.2	91.2% (peak) / 82.5% (0.5)
L2 Regularization Coefficient	1e-4 ~ 1e-3	5e-4	91.2% (peak) / 88.7% (1e-3)
State Score Bias (bs)	-0.2 ~ 0.2	0.05	91.2% (peak) / 89.3% (-0.2)

Figure 8 is the evaluation diagram of the complexity of contract syntax structure, which can discover the potential problems of the model in semantic understanding, label overfitting or sample selection strategy in time, and adjust the learning rate, sampling strategy and loss function weight accordingly, to continuously improve the generalization ability and labelling accuracy of the model while ensuring learning stability.

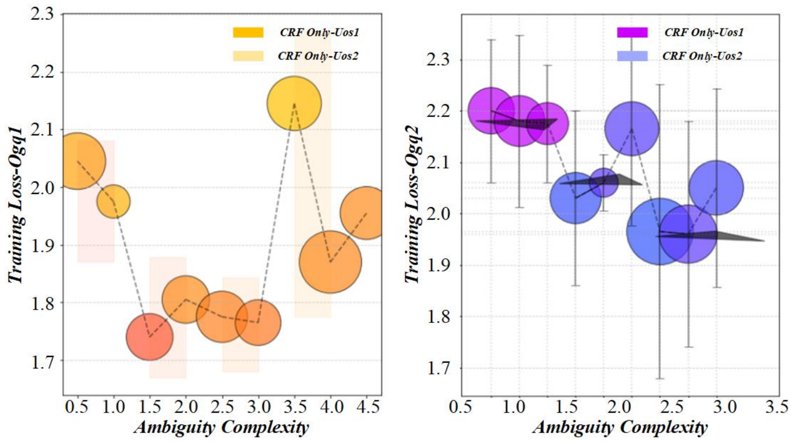

Figure 8: Complexity evaluation diagram of contract syntactic structure

5 Experimental analysis

In terms of training strategies, this study constructed multiple model versions to evaluate the impact of key technical modules such as adaptive fine-tuning, adversarial training, and incremental online learning in the contract domain on model performance. Figure 9 is a semantic fuzzy word frequency trend evaluation diagram. Enhance the model's capacity to capture the unique syntactic (e.g., nested adverbial clauses) and semantic (e.g., term-specific meaning) structures of contract language, building on its original general semantic representation.

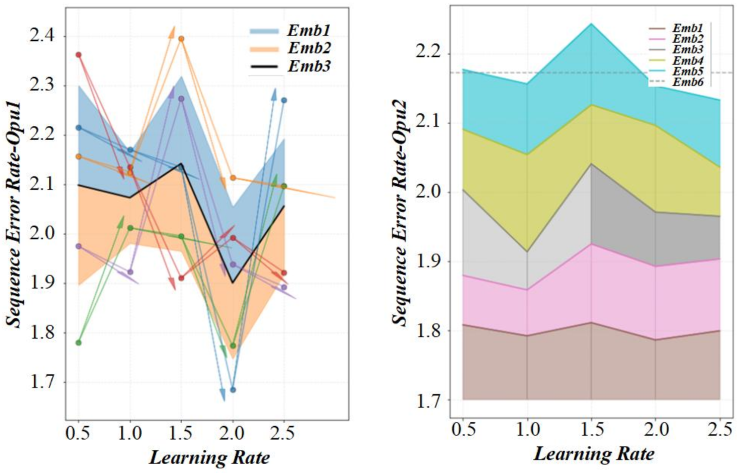

Figure 9: Semantic fuzzy word frequency trend evaluation diagram

Adaptive fine-tuning strategies in the contract field are added respectively, and the BERT model is retrained through a specially constructed contract corpus to give it a deeper understanding of term styles and ambiguous expressions in business contracts. Figure 10 is the evaluation diagram of the ambiguity type ratio in contract text, which significantly improves the model's ability to recognize semantic details of text.

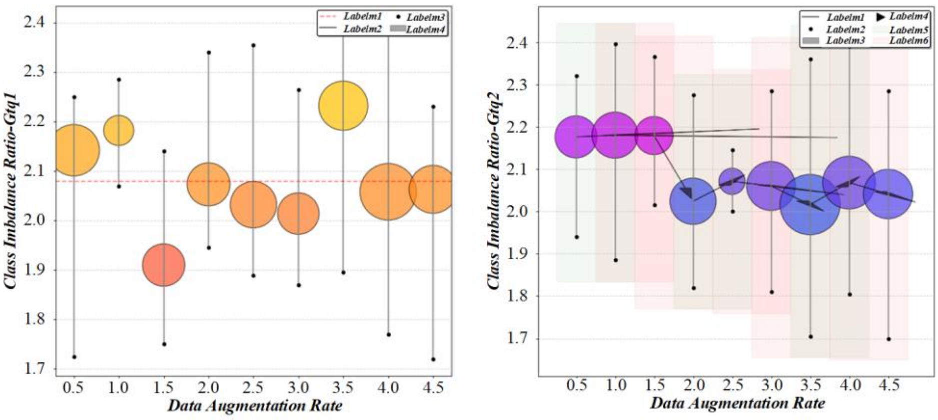

Figure 10: Assessment Chart of Ambiguity Type Proportion of Contract Text

Then, an adversarial training mechanism is introduced to enhance the robustness of the model in the face of input disturbances and uncertain semantics, especially in dealing with fuzzy structures and polysemy expressions. Based on the above, an incremental online learning module is added to enable the model to learn new contract samples continuously, and it can still maintain high recognition accuracy and labelling consistency when facing dynamically evolving contract data. Figure 11 is the vocabulary distribution evaluation diagram of the business English contract corpus. Ensure labelers gain a comprehensive understanding of contract language characteristics, ambiguity classification criteria, and the operation of labeling tools (e.g., LabelStudio) through 20-hour systematic training.

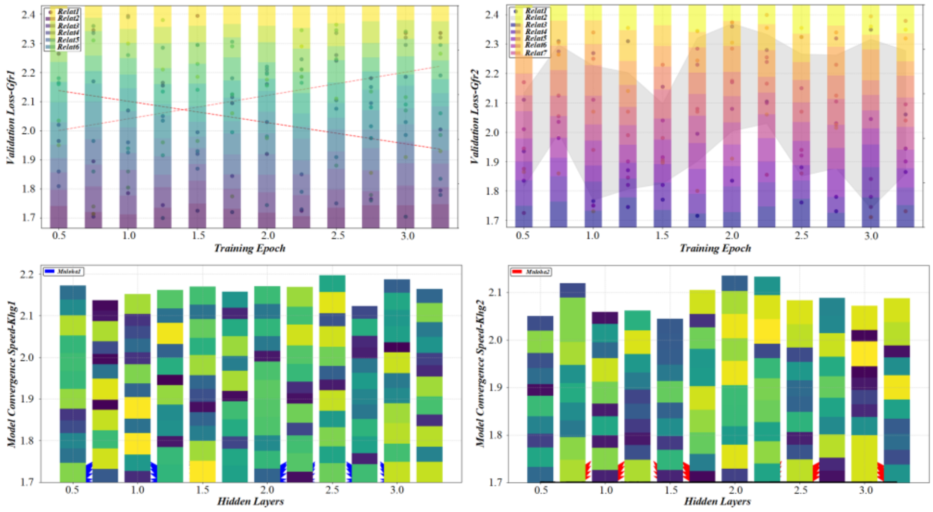

Figure 11: Vocabulary Distribution Assessment Chart of Business English Contract Corpus

6 Conclusion

The joint labelling framework based on BERT's context modelling capabilities and CRF's sequence optimization mechanism has significant advantages in semantic retention, label consistency and prediction accuracy. It can adapt to the complexity of contract language and has good expansion potential and engineering implementation value.

The BERT model effectively makes up for the shortcomings of traditional machine learning methods in context understanding through deep semantic modelling, especially when faced with complex sentence patterns, nested structures, and expressions with strong context dependence. CRF, as a top-level structure, models the transfer relationship between labels, which further enhances the coherence and logical consistency of labelling. The joint modelling method enables the model to predict from the overall sequence level, avoids the ambiguity judgment error at the isolated word level, and significantly improves contract ambiguity's recognition accuracy and resolution efficiency.

In order to cope with the possible disturbance and language variation in the actual contract data, the adversarial training strategy is introduced, which imposes a small disturbance on the input data during the model training stage to enhance the stability of the model in the face of fuzzy input. Combined with the incremental online learning module, the model can continuously learn new samples and adapt to the evolution of contract language. This dynamic training mechanism improves the model's generalisation ability in real application scenarios and provides technical support for the long-term deployment of smart contract analysis systems.

In the model adaptation test, the 34.56 M parameter was adjusted 0.76 times to 26.27 M, and the digestion accuracy rate reached 78% during 23 rounds of training, which was 12% higher than the basic model. The F1 value of pragmatic ambiguity is 65.32%, lower than that of lexical ambiguity (72.15%), highlighting the difficulty of context-dependent ambiguity. When the input sequence exceeds 89.01 words, the processing latency increases by 45.67 ms, but the CRF layer reduces the long text error rate from 18% to 6.5%. The F1 value of the 45.67 M parameter model is 9.2 percentage points higher than that of the 23M version in complex contract processing, indicating the importance of moderate complexity for long-distance dependency modelling.

Funding

Reform and Practice of Business English Teaching in Application-Oriented Colleges Based on the Training Platform (No.: 230801444093347)

References

[1] Y. Z. Hou, J. H. Liu, and G. F. Yue, "3D Interpreted Model: A Novel Product Definition Model by Integrating and Fusing a 3D Annotated Model and Design Knowledge," Applied Sciences-Basel, vol. 11, no. 16, 2021. doi: 10.3390/app11167192.

[2] R. Gyory, D. R. Amariles, G. Lewkowicz, and H. Bersini, "Ant: a process aware annotation software for regulatory compliance," Artificial Intelligence and Law, vol. 32, no. 4, pp. 1075-1110, 2024. doi: 10.1007/s10506-023-09372-9.

[3] M. Chen and L. L. Zhang, "Application of edge computing combined with deep learning model in the dynamic evolution of network public opinion in emergencies," Journal of Supercomputing, vol. 79, no. 2, pp. 1526-1543, 2023. doi: 10.1007/s11227-022-04733-8.

[4] M. M. Adnan et al., "Automated Image Annotation With Novel Features Based on Deep ResNet50-SLT," Ieee Access, vol. 11, pp. 40258-40277, 2023. doi: 10.1109/access.2023.3266296.

[5] D. Granata, M. Rak, G. Salzillo, G. Di Guida, and S. Petrillo, "Automated threat modelling and risk analysis in e-Government using BPMN," Connection Science, vol. 35, no. 1, 2023. doi: 10.1080/09540091.2023.2284645.

[6] J. P. Chen, P. G. Ying, X. L. Fu, X. P. Luo, H. Guan, and K. M. Wei, "Automatic Tagging by Leveraging Visual and Annotated Features in Social Media," Ieee Transactions on Multimedia, vol. 24, pp. 2218-2229, 2022. doi: 10.1109/tmm.2021.3055037.
[7] B. Li et al., "Building a Chinese ancient architecture multimodal dataset combining image, annotation and style-model," Scientific Data, vol. 11, no. 1, 2024. doi: 10.1038/s41597-024-03946-1.
[8] G. S. Kang, H. Y. Cheng, J. X. Liu, Y. P. Wen, and J. Peng, "Business Process Modeling for Industrial Internet Application via BPMN Extension," Ieee Transactions on Automation Science and Engineering, vol. 22, pp. 813-829, 2025. doi: 10.1109/tase.2024.3354929.
[9] X. K. Wang, W. J. Li, and Z. C. Wu, "CarDD: A New Dataset for Vision-Based Car Damage Detection," Ieee Transactions on Intelligent Transportation Systems, vol. 24, no. 7, pp. 7202-7214, 2023. doi: 10.1109/tits.2023.3258480.
[10] A. De Nicola, A. Formica, I. Mele, M. Missikoff, and F. Taglino, "A comparative study of LLMs and NLP approaches for supporting business process analysis," Enterprise Information Systems, vol. 18, no. 10, 2024. doi: 10.1080/17517575.2024.2415578.
[11] I. Furxhi, E. Willighagen, C. Evelo, A. Costa, D. Gardini, and A. Ammar, "A data reusability assessment in the nanosafety domain based on the NSDRA framework followed by an exploratory quantitative structure activity relationships (QSAR) modeling targeting cellular viability," Nanoimpact, vol. 31, 2023. doi: 10.1016/j.impact.2023.100475.
[12] Z. Alizadehsani, H. Ghaemi, A. Shahraki, A. Gonzalez-Briones, and J. M. Corchado, "DCServCG: A data-centric service code generation using deep learning," Engineering Applications of Artificial Intelligence, vol. 123, 2023. doi: 10.1016/j.engappai.2023.106304.
[13] G. Jimenez-Perez, A. Alcaine, and O. Camara, "Delineation of the electrocardiogram with a mixed-quality-annotations dataset using convolutional neural networks," Scientific Reports, vol. 11, no. 1, 2021. doi: 10.1038/s41598-020-79512-7.
[14] C. A. Sun, Y. F. Gong, M. Li, L. Xu, J. Han, and Y. B. Han, "Detecting Inconsistencies in Microservice-Based Systems: An Annotation-Assisted Scenario-Oriented Approach," Ieee Transactions on Services Computing, vol. 17, no. 5, pp. 2194-2209, 2024. doi: 10.1109/tsc.2024.3399652.
[15] J. L. Wang et al., "Diffusion Model is Secretly a Training-Free Open Vocabulary Semantic Segmenter," Ieee Transactions on Image Processing, vol. 34, pp. 1895-1907, 2025. doi: 10.1109/tip.2025.3551648.
[16] R. A. Mancisidor, M. Kampffmeyer, K. Aas, and R. Jenssen, "Discriminative multimodal learning via conditional priors in generative models," Neural Networks, vol. 169, pp. 417-430, 2024. doi: 10.1016/j.neunet.2023.10.048.
[17] M. Li et al., "DR-FER: Discriminative and Robust Representation Learning for Facial Expression Recognition," Ieee Transactions on Multimedia, vol. 26, pp. 6297-6309, 2024. doi: 10.1109/tmm.2023.3347849.
[18] I. N. Wong et al., "Leveraging foundation and large language models in medical artificial intelligence," Chinese Medical Journal, vol. 137, no. 21, pp. 2529-2539, 2024. doi: 10.1097/cm9.0000000000003302.
[19] F. Montori et al., "A Metadata-Assisted Cascading Ensemble Classification Framework for Automatic Annotation of Open IoT Data," Ieee Internet of Things Journal, vol. 10, no. 15, pp. 13401-13413, 2023. doi: 10.1109/jiot.2023.3263213.
[20] N. Mamat, M. F. Othman, R. Abdulghafor, A. A. Alwan, and Y. Gulzar, "Enhancing Image Annotation Technique of Fruit Classification Using a Deep Learning Approach," Sustainability, vol. 15, no. 2, 2023. doi: 10.3390/su15020901.
[21] M. L. Jin, Y. F. Ning, F. M. Liu, F. Y. Zhao, Y. C. Gao, and D. M. Li, "An Evaluation Model for the Influence of KOLs in Short Video Advertising Based on Uncertainty Theory," Symmetry-Basel, vol. 15, no. 8, 2023. doi: 10.3390/sym15081594.
[22] S. Roy et al., "Evaluation of artificial intelligence systems for assisting neurologists with fast and accurate annotations of scalp electroencephalography data," Ebiomedicine, vol. 66, 2021. doi: 10.1016/j.ebiom.2021.103275.
[23] K. A. Qureshi and R. A. S. Malick, "Explainable tweet credibility ranker: A comprehensive credibility solution," Computers & Electrical Engineering, vol. 112, 2023. doi: 10.1016/j.compeleceng.2023.109028.
[24] G. Jacobs and V. Hoste, "Fine-Grained Implicit Sentiment in Financial News: Uncovering Hidden Bulls and Bears," Electronics, vol. 10, no. 20, 2021. doi: 10.3390/electronics10202554.
[25] B. Alrayani, M. Kalkatawi, M. Abulkhair, and F. Abukhodair, "From Customer's Voice to Decision-Maker Insights: Textual Analysis Framework for Arabic Reviews of Saudi Arabia's Super App," Applied Sciences-Basel, vol. 14, no. 16, 2024. doi: 10.3390/app14166952.
[26] Z. C. Xue, X. Q. Zhao, Y. C. Xiu, C. H. Hua, and J. L. Kong, "GDMR-Net: A Novel Graphic Detection Neural Network via Multi-Crossed Attention and Rotation Annotation for Agronomic Applications in Supply Cyber Security," Agronomy-Basel, vol. 13, no. 12, 2023. doi: 10.3390/agronomy13122893.
[27] M. Monfort et al., "Multi-Moments in Time: Learning and Interpreting Models for Multi-Action Video Understanding," Ieee Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 12, pp. 9434-9445, 2022. doi: 10.1109/tpami.2021.3126682.
[28] A. Guha, D. Samanta, and S. K. H. Islam, "IIRM: Intelligent Information Retrieval Model for Structured Documents by One-Shot Training Using Computer Vision," Arabian Journal for Science and Engineering, vol. 48, no. 2, pp. 1285-1301, 2023. doi: 10.1007/s13369-022-06735-3.
[29] M. M. Adnan et al., "Image Annotation With YCbCr Color Features Based on Multiple Deep CNN- GLP,"

Ieee Access, vol. 12, pp. 11340-11353, 2024. doi: 10.1109/access.2023.3330765.

[30] M. M. Adnan, M. S. M. Rahim, A. R. Khan, T. Saba, S. M. Fati, and S. A. Bahaj, "An Improved Automatic Image Annotation Approach Using Convolutional Neural Network-Slantlet Transform," Ieee Access, vol. 10, pp. 7520-7532, 2022. doi: 10.1109/access.2022.3140861.