Automated Entity Extraction for Mergers and Acquisitions Analysis in EDGAR SEC 10-K Filings: A Hybrid Legal-BERT and Semantic Filtering Approach

Antony Mazzarella* Duraimurugan Rajamanickam†

Abstract

Publicly traded companies in the U.S. submit annual 10-K filings to the EDGAR SEC system, providing comprehensive details about financial conditions, business strategies, and corporate actions such as Mergers and Acquisitions (M&A). Traditional methods for analyzing 10-K filings are manual and time-consuming, limiting scalability in corporate due diligence and financial research. This paper introduces a novel automated approach that leverages Legal-BERT for entity extraction, semantic filtering using GloVe embeddings, and contextual filtering with sentence embeddings to identify key entities and provisions related to M&A activities. Our hybrid method achieves a significant improvement in F1-score (up to **87.8%**) compared to using Legal-BERT alone (**78.2%**) by effectively reducing false positives while maintaining high recall. This approach offers a scalable solution for financial analysis, compliance monitoring, risk assessment, and competitive analysis.

1 Introduction

Publicly traded companies in the United States are required to submit 10-K filings to the Securities and Exchange Commission (SEC). These documents provide investors with insights into a company's financial performance and strategic decisions, including mergers, acquisitions, and other corporate restructuring activities. Mergers and Acquisitions (M&A) represent critical financial events that influence market value and company prospects. However,

*University of Arkansas at Little Rock, Little Rock, AR, USA. Email: ajmazzarella@ualr.edu
†University of Arkansas at Little Rock, Little Rock, AR, USA. Email: drajamanicka@ualr.edu

manually identifying and extracting relevant M&A-related information from 10-K filings is labor-intensive and prone to errors.

Recent advancements in Natural Language Processing (NLP), particularly the introduction of transformer-based models like BERT [1] and Legal-BERT [2], have enabled automated text extraction from legal documents. Despite their strong performance in general NLP tasks, models like Legal-BERT often struggle with domain-specific terminology found in 10-K filings, particularly in the context of M&A activities.

This paper proposes a hybrid approach that automates the analysis of 10-K filings, focusing on extracting and filtering entities related to M&A. The innovations in our approach include:

- **Semantic filtering using GloVe embeddings** to capture relevant M&A entities.
- **Contextual filtering using sentence embeddings** to improve entity recognition precision by evaluating text context.

2 Related Work

2.1 Entity Recognition in Legal Texts

Entity extraction from legal and financial documents has advanced significantly with the introduction of transformer-based models like BERT [1]. Legal-BERT, a variant fine-tuned for legal texts, has shown promise in handling legal terminology and complex structures [2]. However, Legal Entity Recognition (LER) in domain-specific areas like M&A still presents challenges due to the ambiguity and variability in legal language. Earlier research, such as the LEDGAR corpus [3], has focused on legal contract classification, providing valuable insights for text classification in legal contexts.

2.2 Semantic Filtering Techniques

Semantic filtering is a post-processing step that enhances the precision of NLP models in specific domains. Previous work using cosine similarity between word embeddings has shown success in filtering out irrelevant entities in specialized domains, including legal and biomedical fields [4]. This paper builds on such techniques by incorporating GloVe embeddings for semantic filtering in M&A-specific contexts.

2.3 Contextual Filtering with Sentence Embeddings

Sentence embeddings, like those derived from Sentence-BERT, have gained popularity for understanding sentence-level context in legal and financial documents [5]. In our approach, contextual filtering using sentence embeddings enables the evaluation of surrounding text, improving the accuracy of entity extraction in complex legal documents such as 10-K filings.

3 Proposed Methodology

This section outlines the data flow, preprocessing steps, model design, and filtering algorithms used in the hybrid approach, which combines Legal-BERT with semantic and contextual filtering to improve the extraction of M&A-related entities from 10-K filings.

3.1 Data Collection from EDGAR SEC

We collected 10-K filings from the EDGAR SEC database using Central Index Key (CIK) and accession numbers of companies involved in recent M&A activities. Our initial focus was on large companies like Apple and IBM due to their prominent M&A activities. To enhance generalizability, we expanded the dataset to include companies from diverse industries such as healthcare, finance, and technology.

Let:

- $C = \{c_1, c_2, \ldots, c_N\}$ be the set of companies.
- For each company c_i, let $F_i = \{f_{i1}, f_{i2}, \ldots, f_{iK}\}$ be the set of their 10-K filings.

3.2 Preprocessing and Section Extraction

We used BeautifulSoup to clean the 10-K filings and remove irrelevant sections, including HTML tags and boilerplate text. Specific sections were targeted based on keywords related to M&A.

Algorithm 1: Preprocessing and Section Extraction

Input:

- Set of filings $\mathcal{F} = \bigcup_{i=1}^{N} F_i$.

- M&A-related keywords $\mathcal{K} = \{k_1, k_2, \ldots, k_M\}$.

Output:

- Extracted sections $\mathcal{S}$.

Procedure:

1. **Initialization:** $\mathcal{S} \leftarrow \emptyset$.
2. **For each filing** $f \in \mathcal{F}$**:**
   (a) **Text Cleaning:**
      - Remove HTML tags: cleaned_text $\leftarrow$ BeautifulSoup(f).

   (b) **Section Identification:**
      - Split cleaned_text into sections $\{s_1, s_2, \ldots, s_L\}$.
      - **For each section** s_j**:**
        – **If** $\exists k \in \mathcal{K}$ such that k appears in s_j:
          * $\mathcal{S} \leftarrow \mathcal{S} \cup \{s_j\}$.
3. **Return** $\mathcal{S}$.

3.3 Entity Recognition with Legal-BERT

After preprocessing, the text was tokenized using the Legal-BERT tokenizer, and Named Entity Recognition (NER) was performed.

Algorithm 2: Entity Recognition with Legal-BERT

Input:

- Extracted sections $\mathcal{S}$.

Output:

- Predicted entities $\mathcal{E}$.

Procedure:

1. **Initialization:** $\mathcal{E} \leftarrow \emptyset$.
2. **For each section** $s \in \mathcal{S}$**:**
   (a) **Tokenization:**
   - Tokens $T = \{t_1, t_2, \ldots, t_n\} \leftarrow \text{Tokenizer}(s)$.

   (b) **NER Prediction:**
   - Labels $L = \{l_1, l_2, \ldots, l_n\} \leftarrow \text{Legal-BERT}(T)$.

   (c) **Entity Extraction:**
   - Extract entities E_s from T and L.
   - $\mathcal{E} \leftarrow \mathcal{E} \cup E_s$.
3. **Return** $\mathcal{E}$.

3.4 Semantic Filtering with GloVe Embeddings

To improve precision, we implemented semantic filtering using GloVe embeddings.

Let:

- $\mathcal{E} = \{e_1, e_2, \ldots, e_P\}$ be the set of extracted entities.
- $\mathcal{P} = \{p_1, p_2, \ldots, p_Q\}$ be predefined M&A-related terms.

For each entity e_i:

1. Compute GloVe embedding $\mathbf{e}_i$.
2. Compute cosine similarity $S(e_i, p_j)$ with each $p_j \in \mathcal{P}$:

$$S(e_i, p_j) = \frac{\mathbf{e}_i \cdot \mathbf{p}_j}{\|\mathbf{e}_i\| \, \|\mathbf{p}_j\|}. \tag{1}$$

3. **Thresholding:** Retain e_i if $\max_j S(e_i, p_j) \geq \theta_s$, where θ_s is the semantic similarity threshold (e.g., 0.8).

Algorithm 3: Semantic Filtering with GloVe Embeddings

Input:

- Entities $\mathcal{E}$, predefined terms $\mathcal{P}$, threshold θ_s.

Output:

- Filtered entities $\mathcal{E}'$.

Procedure:

1. **Initialization:** $\mathcal{E}' \leftarrow \emptyset$.
2. **For each entity** $e_i \in \mathcal{E}$**:**
   (a) **Compute Embedding:** $\mathbf{e}_i \leftarrow \text{GloVe}(e_i)$.
   (b) **For each predefined term** $p_j \in \mathcal{P}$**:**
      - $\mathbf{p}_j \leftarrow \text{GloVe}(p_j)$.
      - Compute $S(e_i, p_j)$.
   (c) **If** $\max_j S(e_i, p_j) \geq \theta_s$:
      - $\mathcal{E}' \leftarrow \mathcal{E}' \cup \{e_i\}$.
3. **Return** $\mathcal{E}'$.

3.5 Contextual Filtering with Sentence Embeddings

Further precision improvements were achieved by applying contextual filtering using Sentence-BERT.

Let:

- $\mathcal{S}_{\mathcal{E}'}$ be sentences containing entities in $\mathcal{E}'$.
- $\mathcal{Q} = \{q_1, q_2, \ldots, q_R\}$ be predefined M&A sentence patterns.

For each sentence $s_k \in \mathcal{S}_{\mathcal{E}'}$:

1. Compute Sentence-BERT embedding $\mathbf{s}_k$.
2. Compute cosine similarity $S(s_k, q_j)$ with each $q_j \in \mathcal{Q}$:

$$S(s_k, q_j) = \frac{\mathbf{s}_k \cdot \mathbf{q}_j}{\|\mathbf{s}_k\| \, \|\mathbf{q}_j\|}. \tag{2}$$

3. **Thresholding:** Retain entities in s_k if $\max_j S(s_k, q_j) \geq \theta_c$, where θ_c is the contextual similarity threshold (e.g., 0.8).

Algorithm 4: Contextual Filtering with Sentence Embeddings

Input:

- Sentences $\mathcal{S}_{\mathcal{E}'}$, patterns $\mathcal{Q}$, threshold θ_c.

Output:

- Final set of entities $\mathcal{E}''$.

Procedure:

1. **Initialization:** $\mathcal{E}'' \leftarrow \emptyset$.
2. **For each sentence** $s_k \in \mathcal{S}_{\mathcal{E}'}$**:**
   (a) **Compute Embedding:** $\mathbf{s}_k \leftarrow \text{Sentence-BERT}(s_k)$.
   (b) **For each pattern** $q_j \in \mathcal{Q}$**:**
      - $\mathbf{q}_j \leftarrow \text{Sentence-BERT}(q_j)$.
      - Compute $S(s_k, q_j)$.
   (c) **If** $\max_j S(s_k, q_j) \geq \theta_c$:
      - $\mathcal{E}'' \leftarrow \mathcal{E}'' \cup \{\text{Entities in } s_k\}$.
3. **Return** $\mathcal{E}''$.

4 Establishing Legal Patterns for Mergers and Acquisitions

To enhance the contextual filtering process, we established a set of legal patterns characteristic of M&A activities.

4.1 Pattern Identification

We analyzed a corpus of legal documents and M&A sections in 10-K filings to identify common linguistic patterns.

- **N-gram Analysis:** Extracted frequent bigrams and trigrams associated with M&A.
- **Dependency Parsing:** Identified syntactic structures prevalent in M&A contexts.

4.2 Pattern Formalization

Patterns $\mathcal{Q}$ were formalized as templates, such as:

- "The Company **entered into** a Merger Agreement with..."
- "We **acquired** all **outstanding shares** of..."
- "On [Date], **completed the acquisition** of..."

Each pattern q_j is represented as a sentence embedding $\mathbf{q}_j$ using Sentence-BERT.

4.3 Incorporation into Contextual Filtering

These patterns serve as benchmarks in Algorithm 4 for evaluating the contextual relevance of sentences containing extracted entities.

5 Experimental Setup

5.1 Dataset

The dataset included 10-K filings from 2020 to 2023 for various companies across industries such as technology, finance, and healthcare. Filings with known M&A activities were selected using CIK and accession numbers. The dataset comprised:

- **Total Companies (N):** 50
- **Total Filings ($\sum_{i=1}^{N} |F_i|$):** 150
- **Total Extracted Sections ($|\mathcal{S}|$):** 1,200

5.2 Baseline Models

We compared four models:

1. **Baseline (Legal-BERT):** Entity extraction using Legal-BERT without filtering.
2. **Model A (Legal-BERT + Semantic Filtering):** Applying Algorithm 3.
3. **Model B (Legal-BERT + Contextual Filtering):** Applying Algorithm 4.
4. **Hybrid Model:** Combining both semantic and contextual filtering (Algorithms 3 and 4).

5.3 Evaluation Metrics

Performance was evaluated using precision (P), recall (R), and F1-score (F_1):

$$P = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad R = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad F_1 = 2 \times \frac{P \times R}{P + R}. \tag{3}$$

- **True Positives (TP):** Correctly identified M&A entities.
- **False Positives (FP):** Incorrectly identified entities.
- **False Negatives (FN):** Missed M&A entities.

6 Performance Evaluation and Ablation Testing

6.1 Ablation Testing

Ablation tests were conducted to assess the contribution of each component in the **Hybrid Model**:

- **Ablation 1 (Hybrid without Semantic Filtering):** Evaluated the impact of removing semantic filtering (Algorithm 3) from the hybrid model.
- **Ablation 2 (Hybrid without Contextual Filtering):** Evaluated the impact of removing contextual filtering (Algorithm 4) from the hybrid model.
- **Baseline (Legal-BERT):** Equivalent to the hybrid model without both filters.

6.2 Results

Table 1: Corrected Performance Metrics

Model	Precision (%)	Recall (%)	F1-score (%)
Baseline (Legal-BERT)	75.0	81.7	78.2
Model A (Semantic Filtering)	82.5	80.2	81.3
Model B (Contextual Filtering)	80.1	83.5	81.8
Hybrid Model	**88.0**	**87.7**	**87.8**
Ablation 1 (Hybrid without Semantic Filtering)	84.5	87.0	85.7
Ablation 2 (Hybrid without Contextual Filtering)	86.2	85.5	85.8

6.3 Discussion

- **Ablation 1 (Hybrid without Semantic Filtering):**
  - Precision decreased to 84.5%, indicating that semantic filtering significantly contributes to reducing false positives.
  - Recall remained relatively high at 87.0%, showing that contextual filtering still captures most relevant entities.
- **Ablation 2 (Hybrid without Contextual Filtering):**
  - Precision decreased to 86.2%, showing that semantic filtering maintains high precision even without contextual information.
  - Recall dropped to 85.5%, highlighting the role of contextual filtering in identifying all relevant entities.
- **Overall Impact:** The hybrid model outperforms all other configurations, confirming the complementary benefits of combining semantic and contextual filtering.

6.4 Discussion

- **Baseline Performance:** The Legal-BERT model achieved an F1-score of 78.2%, indicating decent initial performance but with room for improvement.
- **Semantic Filtering Impact:** Model A improved precision significantly (from 75.0% to 82.5%) by eliminating semantically irrelevant entities, albeit with a slight drop in recall.

Table 2: Performance Metrics

Model	Precision (%)	Recall (%)	F1-score (%)
Baseline (Legal-BERT)	75.0	81.7	78.2
Model A (Semantic Filtering)	82.5	80.2	81.3
Model B (Contextual Filtering)	80.1	83.5	81.8
Hybrid Model	**88.0**	**87.7**	**87.8**
Ablation 1 (Without Semantic)	80.1	83.5	81.8
Ablation 2 (Without Contextual)	82.5	80.2	81.3
Ablation 3 (Without Both Filters)	75.0	81.7	78.2

- **Contextual Filtering Impact:** Model B enhanced recall (from 81.7% to 83.5%) by ensuring entities were contextually relevant, with a modest increase in precision.
- **Hybrid Model Superiority:** Combining both filters resulted in the highest F1-score of 87.8%, demonstrating the complementary benefits of semantic and contextual filtering.

6.5 Statistical Significance

We performed McNemar's test [6] to assess the statistical significance of improvements. The hybrid model's performance gains were statistically significant ($p < 0.01$) compared to the baseline.

7 Results and Discussion

7.1 Error Analysis

- **False Positives Reduced:** Semantic filtering effectively reduced false positives by filtering out entities not semantically related to M&A.
- **False Negatives Addressed:** Contextual filtering recovered some entities that were semantically relevant but initially missed due to atypical phrasing.

7.2 Limitations

- **Dependency on Predefined Terms and Patterns:** The effectiveness of filtering relies on the comprehensiveness of $\mathcal{P}$ and $\mathcal{Q}$.

- **Dynamic Language Usage:** Legal language evolves, and new expressions may not be captured by existing patterns.

7.3 Future Work

- **Adaptive Pattern Learning:** Implementing machine learning techniques to dynamically update $\mathcal{P}$ and $\mathcal{Q}$, such as the methods proposed in [7].
- **Extension to Other Domains:** Applying the hybrid approach to other legal documents and financial filings, potentially incorporating domain adaptation techniques [8].

8 Conclusion

This paper presented a hybrid approach combining Legal-BERT with semantic and contextual filtering to improve entity extraction in M&A analysis from 10-K filings. The incorporation of GloVe embeddings and Sentence-BERT embeddings significantly enhanced precision and recall. The hybrid model achieved an F1-score of 87.8%, outperforming the baseline Legal-BERT model. This approach offers a scalable solution for financial analysis and can be extended to other domains requiring precise entity extraction.

Acknowledgments

We would like to thank the University of Arkansas at Little Rock for supporting this research.

References

[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *NAACL HLT*, 2019, pp. 4171–4186.

[2] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, and I. Androutsopoulos, "LEGAL-BERT: The muppets straight out of law school," in *Findings of EMNLP*, 2020, pp. 2898–2904.

[3] L. Tuggener, R. Otterbach, P. von Däniken, M. Cieliebak, and F. Uzdilli, "LEDGAR: A large-scale multi-label corpus for text classification of legal provisions in contracts," in *LREC*, 2020, pp. 1235–1241.

[4] Y. Kim, H. Lee, and Y. Choi, "Semantic filtered word embedding for improving word similarity," *IEEE Access*, vol. 7, pp. 126 367–126 375, 2019.

[5] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using siamese BERT-networks," in *EMNLP-IJCNLP*, 2019, pp. 3982–3992.

[6] Q. McNemar, "Note on the sampling error of the difference between correlated proportions or percentages," *Psychometrika*, vol. 12, no. 2, pp. 153–157, 1947.

[7] C. Sun, X. Qiu, Y. Xu, and X. Huang, "How to fine-tune BERT for text classification?" in *Chinese Computational Linguistics*, ser. Lecture Notes in Computer Science, vol. 11856. Springer, 2019, pp. 194–206.

[8] S. Gururangan, A. Marasović, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, and N. A. Smith, "Don't stop pretraining: Adapt language models to domains and tasks," in *ACL*, 2020, pp. 8342–8360.