A Multi-Model Approach to Relation Extraction: BERT Embeddings with XGBoost and Graph Neural Networks

Aishwarya Kulkarni (11606062)*, Vaibhav Parihar (11544144)*,
Shreya Abraham Varghese (11603111)*, and Chirag Tolani (14124054)*
*The University of Manchester, Manchester, England, M13 9PL
*{aishwarya.kulkarni, vaibhav.parihar, shreya.varghese, chirag.tolani}@postgrad.manchester.ac.uk

***Abstract*—Relation extraction is crucial for identifying structured relationships between two entities in a given text and has a variety of applications in information extraction. This study mainly compares relation extraction across two ML models: a Graph Convolutional Network (GCN)-based and XGBoost. The first method represents entity relations as a graph using BERT embeddings and graph convolution operations (GCN) to capture dependencies whereas the second method utilizes BERT embeddings (CLS token) and XGBoost in a MultiOutputClassifier framework for multi-label classification, optimizing thresholds via Precision-Recall curves. Both approaches follow a similar method of pre-processing to extract BERT embeddings. While GCN excels at structured relational modeling, XGBoost offers a scalable, interpretable alternative. This paper highlights the strengths and trade-offs between graph-based deep learning and gradient-boosted decision trees models when applied for relation extraction task.**

***Index Terms*—Relation Extraction, BERT Embeddings, XGBoost Classifier, Multi-Label Classification, Graph Convolutional Network**

I. Introduction

Relation extraction is the process of identifying significant relationships between entities in a given text. This is an important task in Natural Language Processing (NLP) and is essential to many real-world applications, such as the creation of knowledge graphs, information retrieval, and question answering. By automatically identifying relations between entities, RE aids in organizing unstructured textual data, enabling machines to understand and process information in a more human-like manner. However, relation extraction remains a challenging task due to the complexities of language, ambiguity in entity interactions, and imbalanced data distributions.

In this study, we tackle these challenges using ReDocRED, a refined version of DocRED dataset, specifically designed to mitigate annotation inconsistencies and improve dataset quality [2]. ReDocRED offers 4053 document-level samples, divided into 3053 for training, 500 for development, and 500 for testing. Each sample includes a "vertexSet" of entities, "labels" specifying relations, and "sents" with tokenized sentences. One striking feature is its imbalance: relations like P131 (20,402 instances) and P17 (14,401 instances) dominate the training set, while rarer ones, such as P1198 and P190, appear exclusively in development or test sets, challenging models to generalize beyond frequent patterns [2]. Entity types follow a similar skew, with locations (LOC) prevalent and numerals (NUM) scarce—only 4149 instances in training—potentially undermining performance on numeric-based relations.

Sentence length adds another layer of difficulty, averaging 25 tokens but ranging from 6 to 180 tokens in training. This variability tests BERT-based models, limited to 512 tokens, where truncation of longer sequences risks losing critical context [4]. Entities typically reside within a single sentence, though some span 13-14 sentences, and while most relations connect entities 2.3 sentences apart on average, some stretch across 16-20 sentences [2]. These traits—imbalanced relations, uneven entity distributions, variable sentence lengths, and long-range dependencies—highlight the limitations of sentence-level methods and demand a document-level approach.

To meet these demands, we employ BERT for rich contextual embeddings, paired with XGBoost and a Graph Convolutional Network (GCN) to capture both semantic depth and structural relationships across ReDocRED's diverse documents. This dual strategy aims to robustly extract relations despite the dataset's inherent complexities.

II. LITERATURE REVIEW

Recent advancements in relation extraction (RE) have significantly benefited from transformer-based models, particularly BERT, due to their exceptional ability to encode rich contextual information from unstructured texts. Zhang et al. introduced BERT-XML, pre-trained specifically on medical-domain texts, effectively handling specialized vocabulary through multilabel attention mechanisms[5]. This approach has notably improved performance in large-scale automated ICD coding tasks by leveraging domain-specific contexts embedded in the medical literature [7]. The superiority of BERT-based embeddings lies primarily in their ability to dynamically adapt contextual information, enabling a deeper understanding of nuanced semantic interactions between entities in a given textual corpus.

Ensemble methods, especially gradient-boosting frameworks such as XGBoost, have also gained popularity in NLP due to their interpretability, strong predictive performance, and computational efficiency. XGBoost, an ensemble learning

algorithm based on gradient boosting, effectively manages structured and semi-structured datasets, excelling particularly when data are noisy or sparse. Its built-in regularization techniques, such as L1 and L2 penalties, and early stopping capabilities further help reduce the risk of overfitting, enhancing generalization on complex tasks. The integration of XGBoost with BERT-derived embeddings has demonstrated substantial improvements in predictive accuracy while maintaining simplicity in model architecture, which requires fewer hyperparameter tuning efforts compared to deep learning methods [6].

In parallel, Graph Convolutional Networks (GCNs) have demonstrated their strength in handling structured relational information embedded within documents. Zhou et al. developed Global Context-enhanced Graph Convolutional Networks (GCGCN), effectively addressing the complexities of document-level relation extraction through hierarchical inference blocks and context-aware attention mechanisms. GCGCN progressively integrates context information, capturing both local (sentence level) and global (document level) interactions among entities, achieving state-of-the-art performance in tasks demanding multi-hop reasoning [1]. Further highlighting improvements in relation extraction, Guo et al. highlighted the critical role of attention mechanisms in enhancing GCN performance by incorporating dynamic attention to selectively emphasize relevant graph substructures, significantly increasing the capability of GCNs to detect and interpret relational interactions within complex textual datasets [3].

Finally, annotation quality has been a pivotal aspect influencing model performance in relation extraction tasks. Recognizing issues such as false-negative annotations within the DocRED dataset, Tan et al. proposed Re-DocRED, a meticulously re-annotated dataset that resolves these annotation inaccuracies. This enhanced dataset substantially improves the quality of training and evaluation, enabling models to achieve greater precision and recall. The refinement in dataset annotation illustrates the essential role accurate data plays in advancing relation extraction research and improving practical applicability [2].

III. APPROACH AND METHODOLOGY

A. Preprocessing with BERT

BERT (Bidirectional Encoder Representations from Transformers) serves as the foundational preprocessing tool, transforming raw text into high-quality contextual embeddings. Unlike traditional word embeddings, the bidirectional attention mechanism of BERT allows it to capture dynamic dependencies between words, which is crucial to understanding nuanced semantic relationships [4]. The preprocessing pipeline begins with WordPiece tokenization, which decomposes words into subword units, enabling the model to handle rare and out-of-vocabulary words efficiently. Each token is encoded using three distinct embedding types: token embeddings (word representation), segment embeddings (sentence distinction), and position embeddings (word order). These embeddings are processed through multiple layers of transformer encoders, utilizing self-attention mechanisms, assigning importance weights dynamically to words in a sentence. The attention mechanism is computed as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

where Q, K, V are query, key, and value matrices derived from the input embeddings. This enables BERT to retain syntactic and semantic relationships across long textual sequences [4].

The preprocessing step involves loading training, development, and test datasets, followed by relation filtering to remove labels present only in evaluation datasets. Relations are then encoded into multi-label binary vectors using `MultiLabelBinarizer`, and feature extraction is performed using `bert-base-uncased`. Tokenized sentences are then truncated or padded to fit a 512-token limit, and the `CLS` token embedding is then extracted as the final sentence representation.

B. Classification with XGBoost

XGBoost (eXtreme Gradient Boosting) is employed as the primary classification model due to its efficiency in handling structured data and multi-label classification. Unlike traditional boosting methods, XGBoost integrates L1 (Lasso) and L2 (Ridge) regularization, preventing overfitting and improving generalization [6]. It also effectively processes missing values and sparse data by learning optimal split directions dynamically. In our implementation, a MultiOutputClassifier framework is utilized to train separate binary classifiers for each relation type. The model's hyperparameters are optimized with 700 boosting iterations, a maximum tree depth of 6, and 80% feature sampling from BERT embeddings to enhance speed and reduce overfitting.

Class imbalance is addressed by calculating `scale_pos_weights` as:

$$\frac{(N - \text{class\_count})}{(\text{class\_count} + 10^{-6})}$$

Performance is further refined by tuning decision thresholds using precision-recall curves on the development set, ensuring balanced precision and recall across all relation types.

C. Graph-Based Learning with GCN

Graph Convolutional Networks (GCNs) are utilized to capture document-level relational structures, extending traditional convolutional operations to graph-structured data[1]. Unlike conventional methods that rely on word proximity, GCNs aggregate contextual information from neighboring entities, allowing for robust multi-hop reasoning.

In our model, BERT embeddings serve as node features, while edges are formed using k-nearest neighbors ($k = 5$) through `knn_graph`, which are then converted into an undirected structure for bidirectional message passing. The GCN architecture comprises of two `GCNConv` layers, that reduces the embedding dimension from 768 to 256, are then followed

by a classification layer. Training employs MultiLabelFocalLoss, which is calculated as:

$$\text{Focal Loss} = -\alpha(1 - p_t)^{\gamma} \log p_t$$

where $\gamma = 2.0$, modulating the focus on hard-to-classify samples. The Adam optimizer (learning rate = 0.005) is used for training over 50 epochs, with loss monitoring and GPU memory optimization.

This combined approach integrates deep contextual embeddings, gradient-boosted decision trees, and graph-based learning to achieve high accuracy in document-level relation extraction.

IV. PERFORMANCE EVALUATION

The two models on RE-DocRED are assessed with the following key metrics:

- **Micro F1:** It combines all labels' true positives, false positives, and false negatives into a single F1 score, reflecting overall performance amid imbalance.
- **Weighted F1:** It averages per-label F1 scores, weighted by support, balancing frequent and rare relations.
- **Micro Ign F1:** It computes Micro F1 on positive predictions only, highlighting true triple detection (likely ignoring negatives).
- **Weighted Ign F1:** It averages positive-label F1 scores by support, detailing performance across frequencies (assumed weighted, ignoring negatives).
- **Precision and Recall:** Precision ensures predicted relations are reliable, while Recall captures most actual relations, both refined by threshold tuning.

Together, these metrics align with measuring performance on extracting meaningful relations from the mentioned two approaches.

V. COMPARATIVE ANALYSIS

The comparative analysis of XGBoost and GCN on the RE-DocRED dataset reveals distinct performance trends across the Dev and Test sets. On the Dev set, XGBoost achieved a Micro F1 of 0.6605 and a Weighted F1 of 0.6708, slightly outperforming GCN's Micro F1 of 0.6013 and Weighted F1 of 0.6496, indicating XGBoost's better overall balance in predicting relations. However, GCN showed a higher Ign F1 (Weighted: 0.6496, Micro: 0.6013) compared to XGBoost's Ign F1 (Weighted: 0.6708, Micro: 0.6605), though the difference is less pronounced when ignoring negatives. XGBoost excelled in Recall at 0.6743 versus GCN's 0.7211, suggesting GCN captures more true relations, but XGBoost's Precision of 0.4997 surpassed GCN's 0.5157, indicating fewer false positives. On the Dev set, XGBoost generally demonstrates a slight edge in balanced performance, while GCN prioritizes recall.

On the Test set, XGBoost maintained its lead with a Micro F1 of 0.6400 and Weighted F1 of 0.6555, compared to GCN's Micro F1 of 0.5546 and Weighted F1 of 0.6046, highlighting XGBoost's superior generalization to unseen data. The Ign F1 metrics further support this, with XGBoost at 0.6400 (Micro) and 0.6555 (Weighted) against GCN's 0.5546 (Micro) and 0.6046 (Weighted), showing XGBoost's consistency in positive relation prediction. Precision and Recall on the Test set also favor XGBoost, with a Precision of 0.4516 and Recall of 0.6572, compared to GCN's Precision of 0.4687 and Recall of 0.6789; although GCN slightly edges out in Recall, XGBoost's higher Precision reflects better reliability in predictions. Overall, XGBoost consistently outperforms GCN across most metrics on both sets, particularly in balanced F1 scores and precision, while GCN shows strength in recall, capturing more true relations at the cost of more false positives.

TABLE I
PERFORMANCE METRICS FOR DEV DATASET

Model	F1		Ign F1		Precision	Recall
	Weighted	Micro	Weighted	Micro		
XGBOOST	0.6708	0.6605	0.6708	0.6605	0.4997	0.6743
GCN	0.6496	0.6013	0.6496	0.6013	0.5157	0.7211

TABLE II
PERFORMANCE METRICS FOR TEST DATASET

Model	F1		Ign F1		Precision	Recall
	Weighted	Micro	Weighted	Micro		
XGBOOST	0.6555	0.6400	0.6555	0.6400	0.4516	0.6572
GCN	0.6046	0.5546	0.6046	0.5546	0.4687	0.6789

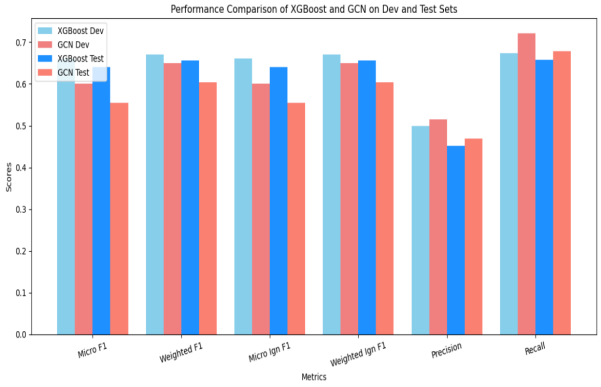

Fig. 1. Performance Metrics Comparison for XGBOOST and GCN

A. XGBoost

- **Strengths:** Demonstrates a robust ability to balance precision and recall, offering consistent performance across datasets, which highlights its adaptability and reliability in predicting relations. Its focus on minimizing false positives enhances its suitability for applications where accuracy is critical, while its stability in handling positive predictions makes it a versatile choice.
- **Weaknesses:** Tends to miss some true relations due to a relatively lower emphasis on recall, which could be a limitation in scenarios where capturing all possible relations is a priority, potentially requiring complementary methods to address this gap.

B. GCN

- **Strengths:** Excels at capturing a broader range of true relations, making it particularly effective in tasks where completeness is essential, such as comprehensive relation extraction. Its strength in specific positive predictions suggests potential advantages in nuanced scenarios where negative instances are less relevant.
- **Weaknesses:** Struggles with a higher rate of false positives due to lower precision, which may undermine reliability, and its overall performance balance is less consistent across datasets, indicating challenges in generalizing to new data. Lower Precision (0.5157 Dev, 0.4687 Test) results in more false positives, reducing prediction reliability.

VI. LIMITATIONS

The following are some of the limitations that arised during our comparative study:

- We utilized `bert-base-uncased` for extracting embeddings for both the approaches instead of larger BERT models due to system limitations.
- Given the limitations of our current system, training time for both the models were extremely high and hence the performance metrics provided might not be the best accurate representation of the entity pairs.
- While we utilized a smaller BERT model for extraction of embeddings, we considered sentence embeddings over word embeddings as training time with word embeddings increased by a significant amount.

VII. CONCLUSION

In conclusion, the evaluation of the GCN and XGBoost models on the RE-DocRED dataset reveals that XGBoost outperformed GCN, achieving a Micro F1 of 0.6605 compared to GCN's 0.5546, with stronger results in Weighted F1, Precision, and Recall. This highlights XGBoost's effectiveness, bolstered by class weighting, in multi-label relation prediction. Despite limitations such as handling rare relations and computational demands, both models demonstrated robust performance overall, successfully predicting relations and contributing to the advancement of machine learning in relation extraction tasks.

REFERENCES

[1] Huiwei Zhou, Yibin Xu, Weihong Yao, Zhe Liu, Chengkun Lang, and Haibin Jiang. 2020. Global Context-enhanced Graph Convolutional Networks for Document-level Relation Extraction. In Proceedings of the 28th International Conference on Computational Linguistics, pages 5259–5270, Barcelona, Spain (Online). International Committee on Computational Linguistics.

[2] (Tan et al., EMNLP 2022)[Revisiting DocRED - Addressing the False Negative Problem in Relation Extraction](https://aclanthology.org/2022.emnlp-main.580/)

[3] (Guo et al., ACL 2019)Zhijiang Guo, Yan Zhang, and Wei Lu. 2019. Attention Guided Graph Convolutional Networks for Relation Extraction. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 241–251, Florence, Italy. Association for Computational Linguistics.

[4] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding.

[5] Revisiting DocRED – Addressing the False Negative Problem in Relation Extraction Qingyu Tan* 1, 2 Lu Xu* 1, 3 Lidong Bing† 1 Hwee Tou Ng2 Sharifah Mahani Aljunied1 1DAMO Academy, Alibaba Group 2Department of Computer Science, National University of Singapore 3Singapore University of Technology and Design

[6] https://en.wikipedia.org/wiki/XGBoost

[7] Zachariah Zhang, Jingshu Liu, and Narges Razavian. 2020. BERT-XML: Large Scale Automated ICD Coding Using BERT Pretraining. In Proceedings of the 3rd Clinical Natural Language Processing Workshop, pages 24–34, Online. Association for Computational Linguistics.