Text Mining Final Assignment - Comparative Analysis of Transformer Models for ADE Extraction on the CADEC Corpus

Emphasizing Span-Level Annotation, and Preprocessing Strategies

Monish Shah (S4401379)

February 5, 2025

1 Introduction

Adverse drug events (ADEs) often surface only after a drug reaches the market, posing serious risks to patient safety and imposing high costs on healthcare systems. Although clinical trials provide initial safety data, they may not reveal all potential adverse effects. Consumer-generated data from platforms such as the AskaPatient forum now offer a rich, real-world source for ADE detection; however, these texts are noisy, and annotations are often discontinuous or split into multiple spans. Such fragmentation can lead to an incomplete capture of the ADE, thereby reducing the effectiveness of downstream extraction models.

In this study, I address these challenges by investigating:

RQ1: How do transformer models pretrained on general-domain (BERT) versus biomedical/clinical corpora (BioBERT and ClinicalBERT) perform for ADE extraction from noisy consumer-generated text?

RQ2: How does the choice of annotation scheme (Original, MedDRA, SCT) affect the ability of models to capture complete ADE spans?

RQ3: What is the impact of different preprocessing strategies (split vs. merge for multi-span annotations) on token-label alignment and downstream performance?

RQ4: Can the architecture of SpanBERT, even in its baseline (un-finetuned) form, provide a robust solution for span-level ADE extraction?

In the sections that follow, I review relevant prior work, detail my data and methodology (with a special focus on span-level precision), present my experimental results and error analysis, and conclude with insights and suggestions for future work.

2 Background and Related Work

Early approaches to ADE extraction predominantly relied on lexicon-based and rule-based methods [6], which were limited in their ability to handle the linguistic variability of consumer-generated text. The CADEC corpus, introduced by Karimi et al. (2015) [1], provided a large-scale annotated dataset for ADE extraction from social media; however, it also revealed challenges such as discontinuous spans and low interrater agreement.

The advent of transformer-based models has revolutionized natural language processing. Devlin et al. (2019) [2] introduced BERT, which set new standards with its deep bidirectional representations. Building on this, Lee et al. (2020) [3] and Alsentzer et al. (2019) [4] developed BioBERT and ClinicalBERT, respectively, which are tailored for biomedical text mining. Joshi et al. (2020) [5] further advanced this area with SpanBERT, a model designed to enhance span-level representations—a critical feature when dealing with discontinuous or multi-part annotations.

Recent studies have reinforced the need for robust preprocessing strategies. Scaboro et al. (2023) [7] conducted an extensive evaluation of transformer architectures for ADE extraction from informal texts and underscored the importance of model selection and proper handling of discontinuous spans. Similarly, Dong et al. (2024) [8] demonstrated that BERT-based models can achieve high accuracy in extracting ADEs from social media data. Other works [9, 10, 11] have highlighted that incomplete span capture degrades performance, supporting my decision to explore both split and merge preprocessing strategies. These findings collectively justify my use of multiple transformer models (including a baseline un-finetuned SpanBERT and RoBERTa) and my emphasis on accurate span-level extraction.

3 Data

The CADEC corpus [1] consists of 1250 posts extracted from the AskaPatient forum. Each post is accompanied by an annotation file (with the extension `.ann`) and is available in three distinct annotation schemes:

- **Original:** Human-annotated spans capturing entities such as `ADR` (adverse drug reactions), `Drug`, `Disease`, `Symptom`, and `Finding`.
- **MedDRA:** Annotations normalized to the Medical Dictionary for Regulatory Activities, facilitating regulatory consistency.
- **SCT:** Annotations mapped to SNOMED Clinical Terms (SCT), which help bridge the gap between layperson language and formal clinical terminology.

For example, a sample from the original annotations is:

```
T1    ADR 32 41    headache
T2    ADR 0 21     nausea
```

A corresponding MedDRA sample is:

```
T1    10040617 32 41    headache
T2    10003068 0 21     nausea
```

And a corresponding SCT sample is:

```
TT1   277521002 | Loss of appetite | 32 41    headache
TT2   CONCEPT_LESS 0 21       nausea
```

For consistency, all annotations are remapped to a binary label (`"ADR"`). In addition, I experiment with two preprocessing strategies for handling multi-span annotations:

- **Split:** Each annotation span is treated independently. This preserves the original segmentation but may result in fragmented ADEs.
- **Merge:** Multiple spans belonging to the same annotation are merged into one continuous span by taking the minimum start and maximum end. This strategy helps capture the complete context of the ADE, which is critical as incomplete spans can lead to suboptimal extraction performance [11].

Statistical Overview and Visualizations

statistics for the corpus:

- **Original:** 1250 documents, a total of 9111 annotations (average of 7.29 annotations per document), with an average span length of 13.57 characters.
- **MedDRA:** 1250 documents, a total of 6318 annotations (average of 5.05 annotations per document), with an average span length of 15.05 characters.
- **SCT:** 1250 documents, a total of 9111 annotations (average of 7.29 annotations per document), with an average span length of 13.30 characters.

Figure 1 shows a boxplot of the number of annotations per document for each annotation scheme. As seen in the figure, the Original and SCT schemes yield higher annotation counts per document than the MedDRA scheme. Figure 2 illustrates the histogram of annotation span lengths, highlighting a wide variation due to the presence of discontinuous spans. Finally, Figure 4 displays a histogram of document word counts, indicating that document lengths vary significantly. These visualizations are critical for understanding the challenges in token-level alignment and the need for an effective merge strategy.

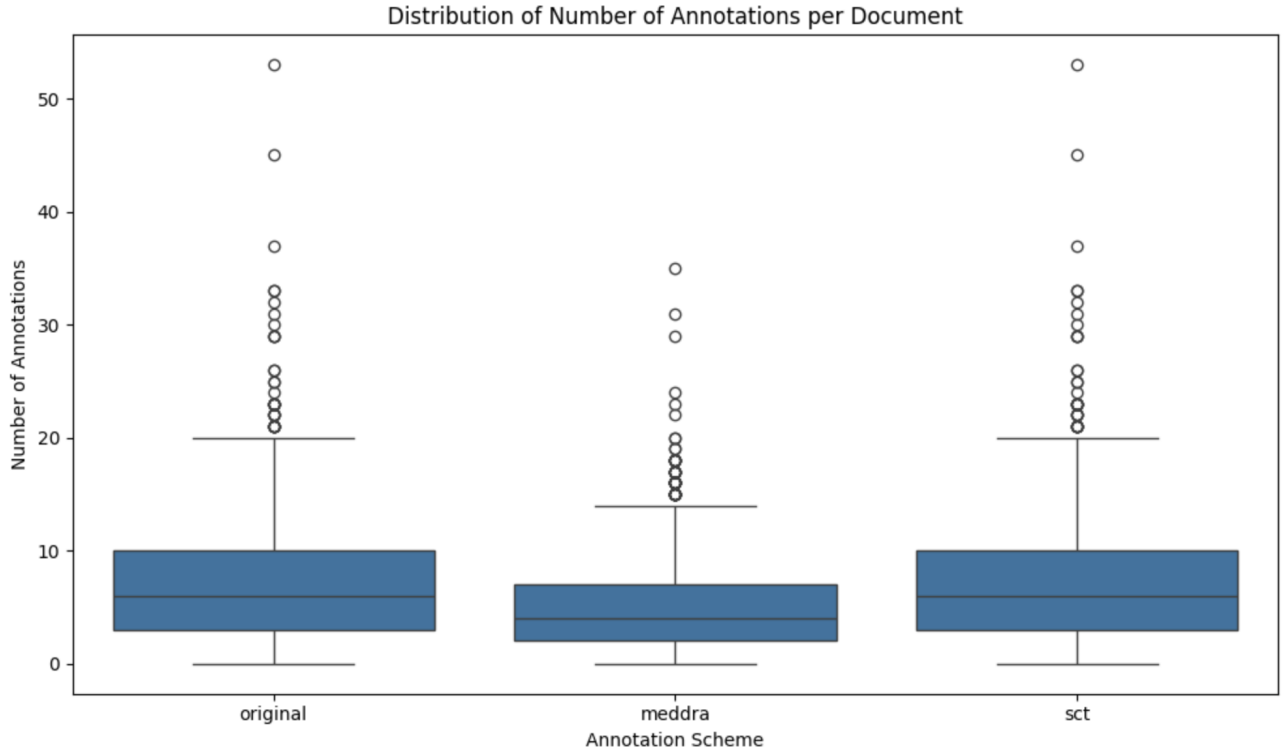

Figure 1: Boxplot of the number of annotations per document for the Original, MedDRA, and SCT annotation schemes. This plot shows that the Original and SCT schemes have higher and more variable annotation counts compared to MedDRA.

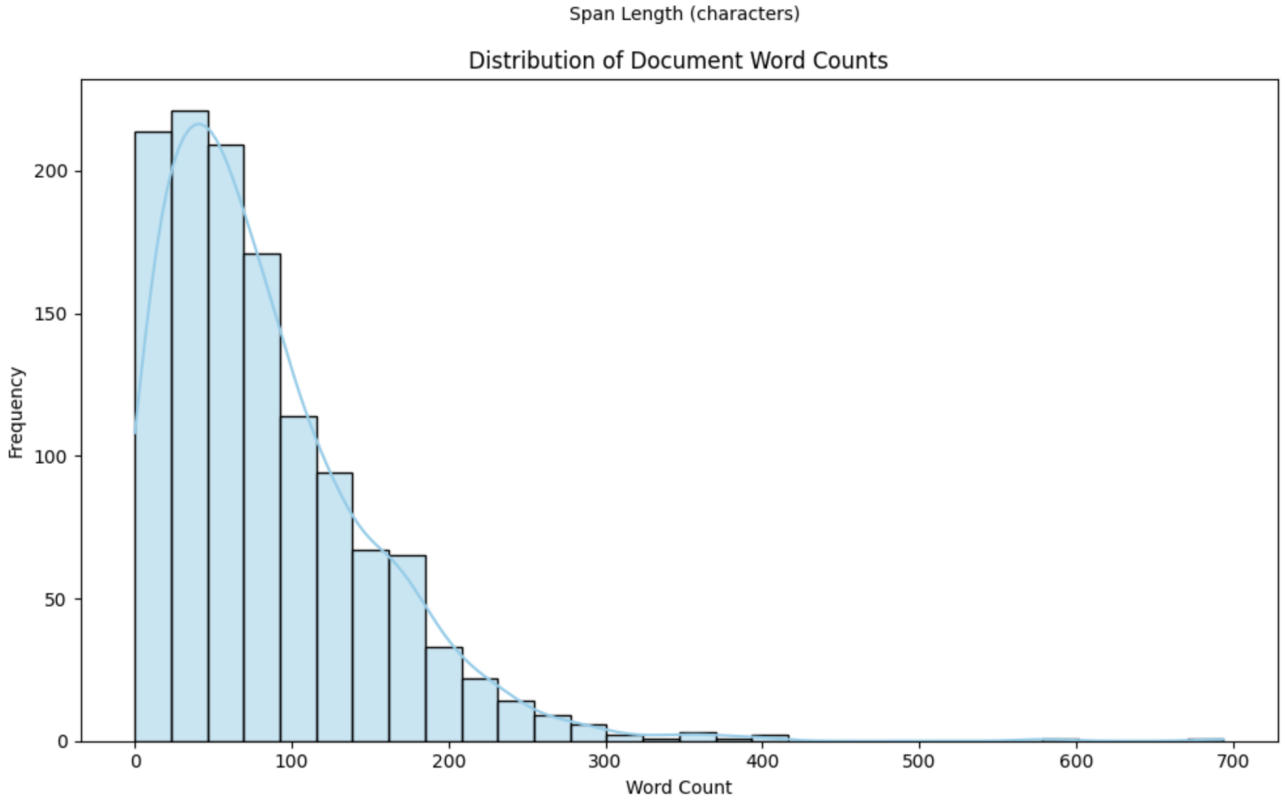

Figure 2: Histogram of annotation span lengths (in characters) for the three annotation schemes. The wide variation underscores the challenges of handling discontinuous spans.

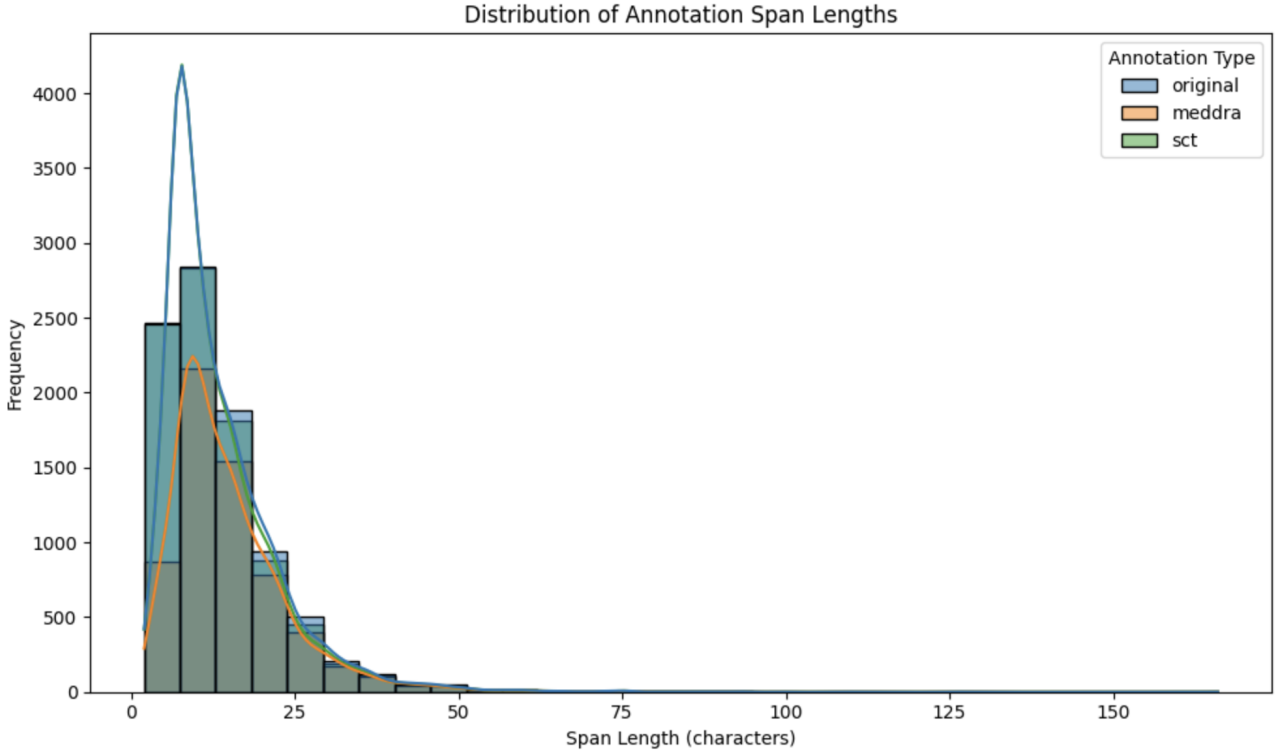

Figure 3: Histogram of document word counts in the CADEC corpus, revealing substantial variation in post lengths.

4 Methods

4.1 Data Parsing and Preprocessing

I developed a parsing pipeline that:

1. Reads the corresponding `.txt` and `.ann` files from three folders (Original, MedDRA, SCT).
2. Processes multi-span annotations using either a **split** strategy (implemented in `parse_and_split_ann_line`) or a **merge** strategy (implemented in `merge_ann_line`).
3. Remaps all labels to `"ADR"` for standardized comparisons.

The function `parse_brat_files_strategy` allows the selection of the desired preprocessing strategy. The merge strategy is particularly critical for consolidating discontinuous spans to ensure full ADE coverage.

4.2 Tokenization and Label Alignment

Using Hugging Face's fast tokenizers, I convert each document's text into tokens. Simultaneously, I create a character-level label array from the parsed annotations. I then use the tokenizer's offset mapping to align these character-level labels to the corresponding tokens. This process is crucial because it ensures that the model receives accurate token-level supervision that reflects the complete spans of ADEs—even when annotations are discontinuous.

4.3 Model Fine-Tuning

I fine-tune several transformer models on the SCT data using both the split and merge preprocessing strategies. my model selection is based on prior work and my goal of accurately capturing spans:

- **BERT** (`bert-base-uncased`): A general-domain transformer model that serves as a baseline.
- **BioBERT** (`dmis-lab/biobert-v1.1`): Pretrained on biomedical literature, this model is designed to better handle clinical terminology.
- **ClinicalBERT** (`emilyalsentzer/Bio_ClinicalBERT`): Adapted for clinical text, it further refines the representations for healthcare-related tasks.
- **SpanBERT (finetuned)** (`abhibisht89/spanbert-large-cased-finetuned-ade_corpus_v2`): Finetuned on an ADE corpus, this model is optimized for span-level tasks.

- **SpanBERT (baseline)** (`SpanBERT/spanbert-large-cased`): The un-finetuned version, used as a benchmark to measure the effect of finetuning on span extraction.
- **RoBERTa** (`roberta-base`): A robust model known for its strong performance on a variety of NLP tasks.

For each model, I use a train/dev/test split of approximately 70%/15%/15%. For models with classifier dimension mismatches (e.g., the SpanBERT variants), I set `ignore_mismatched_sizes=True`.

4.4 Quantitative Analysis

my evaluation framework includes:

1. A summary table of evaluation metrics (precision, recall, F1, and loss) for each model and preprocessing strategy.
2. Comparative analyses using bar plots of Test F1 scores.
3. Analysis of label frequency distributions.

4.5 Error Analysis

I perform error analysis by:

- Generating token-level confusion matrices to highlight trends in misclassification.

This qualitative analysis reveals that the merge strategy more reliably captures the full span of ADEs, particularly at annotation boundaries.

5 Results

In my experiments I conducted evaluations on three annotation schemes (Original, MedDRA, and SCT) using two preprocessing strategies (split and merge) on a total of 1250 annotated documents per scheme. For each setting, I fine-tuned several transformer-based models and computed evaluation metrics (loss, precision, recall, and F1) on both the development and test sets at epoch 3.

Below I summarize the key quantitative findings and then provide an in-depth discussion of the error analysis results.

5.1 Quantitative Results

5.1.1 Original Annotations

Table 1: Evaluation Metrics on Original Annotations (Split and Merge Strategies)

Model (Strategy)	Dev				Test			
	Loss	Precision	Recall	F1	Loss	Precision	Recall	F1
BERT (split)	0.2046	0.6849	0.7329	0.7081	0.1849	0.6760	0.7455	0.7091
BioBERT (split)	0.2226	0.7095	0.7681	0.7376	0.2226	0.6822	0.7569	0.7176
ClinicalBERT (split)	0.1979	0.7084	0.7463	0.7269	0.1917	0.6833	0.7448	0.7128
SpanBERT_finetuned (split)	0.2025	0.7252	0.7556	0.7401	0.1800	0.7290	0.7754	0.7515
SpanBERT_normal (split)	0.1963	0.7222	0.7704	0.7455	0.1842	0.7137	0.7762	0.7437
RoBERTa (split)	0.1953	0.7212	0.7455	0.7332	0.1838	0.6874	0.7457	0.7154
BERT (merge)	0.2274	0.6911	0.7553	0.7218	0.2070	0.6941	0.7701	0.7302
BioBERT (merge)	0.2213	0.7295	0.7899	0.7585	0.2213	0.6880	0.7595	0.7220
ClinicalBERT (merge)	0.1982	0.7279	0.7723	0.7495	0.1905	0.6934	0.7525	0.7218
SpanBERT_finetuned (merge)	**0.1958**	**0.7694**	**0.7852**	**0.7772**	**0.1661**	**0.7623**	**0.7699**	**0.7661**
SpanBERT_normal (merge)	0.1940	0.7397	0.7887	0.7634	0.1752	0.7277	0.7644	0.7456
RoBERTa (merge)	0.1937	0.7239	0.7704	0.7464	0.1837	0.7088	0.7697	0.7380

Table 2: Evaluation Metrics on Meddra Annotations (Split and Merge Strategies)

Model (Strategy)	Dev				Test			
	Loss	Precision	Recall	F1	Loss	Precision	Recall	F1
BERT (split)	0.2071	0.5916	0.6674	0.6272	0.1786	0.6023	0.7000	0.6475
BioBERT (split)	0.2298	0.6035	0.6794	0.6392	0.1919	0.6125	0.6967	0.6519
ClinicalBERT (split)	0.2089	0.5567	0.6920	0.6170	0.1735	0.6078	0.7167	0.6578
SpanBERT_finetuned (split)	0.2147	0.5802	0.6670	0.6206	0.1782	0.6550	0.7312	0.6910
SpanBERT_normal (split)	0.2150	0.5550	0.6182	0.5849	0.1913	0.6190	0.7050	0.6337
RoBERTa (split)	0.2106	0.5320	0.6539	0.5867	0.1737	0.6260	0.7154	0.6677
BERT (merge)	0.2183	0.5793	0.6752	0.6236	0.1832	0.5801	0.6839	0.6277
BioBERT (merge)	0.2472	0.6028	0.7042	0.6495	0.2123	0.5885	0.6931	0.6365
ClinicalBERT (merge)	0.2348	0.5977	0.6936	0.6421	0.1974	0.6103	0.7056	0.6545
SpanBERT_finetuned (merge)	**0.2094**	**0.6589**	**0.7267**	**0.6911**	**0.1710**	**0.6697**	**0.7360**	**0.7013**
SpanBERT_normal (merge)	0.2175	0.6525	0.7132	0.6815	0.1766	0.6389	0.7094	0.6723
RoBERTa (merge)	0.2001	0.6147	0.7094	0.6587	0.1773	0.6404	0.7215	0.6786

Table 3: Evaluation Metrics on SCT Annotations (Split and Merge Strategies)

Model (Strategy)	Dev				Test			
	Loss	Precision	Recall	F1	Loss	Precision	Recall	F1
BERT (split)	0.2315	0.6662	0.7348	0.6988	0.2113	0.6642	0.7515	0.7051
BioBERT (split)	0.2345	0.6875	0.7643	0.7238	0.2279	0.6702	0.7668	0.7153
ClinicalBERT (split)	0.2065	0.6887	0.7536	0.7197	0.1996	0.6657	0.7491	0.7049
SpanBERT_finetuned (split)	0.2172	0.7081	0.7657	0.7358	0.1904	0.7150	0.7705	0.7417
SpanBERT_normal (split)	0.2134	0.7058	0.7593	0.7316	0.1925	0.7095	0.7730	0.7399
RoBERTa (split)	0.2085	0.6813	0.7327	0.7061	0.1943	0.6647	0.7517	0.7055
BERT (merge)	0.2370	0.6755	0.7566	0.7138	0.2165	0.6793	0.7669	0.7205
BioBERT (merge)	0.2164	0.6929	0.7816	0.7345	0.2083	0.6762	0.7626	0.7168
ClinicalBERT (merge)	0.2091	0.7040	0.7705	0.7358	0.2030	0.6823	0.7535	0.7161
SpanBERT_finetuned (merge)	**0.2042**	**0.7379**	**0.7926**	**0.7643**	**0.1829**	**0.7343**	**0.7827**	**0.7577**
SpanBERT_normal (merge)	0.2134	0.7270	0.7829	0.7539	0.1945	0.7234	0.7752	0.7484
RoBERTa (merge)	0.2061	0.7077	0.7776	0.7410	0.1929	0.7016	0.7720	0.7351

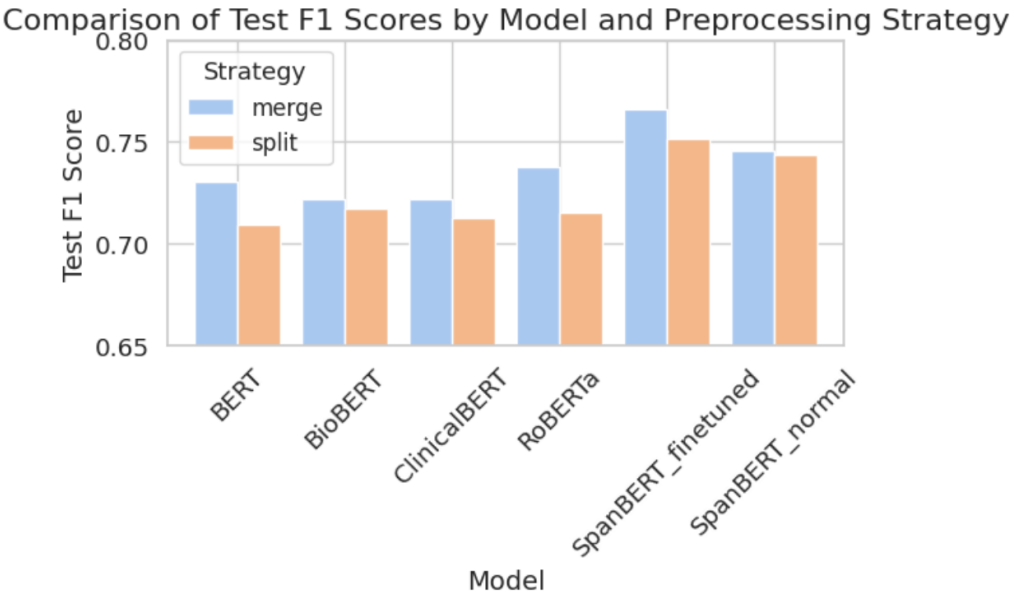

Figure 4: Histogram of document word counts in the CADEC corpus, revealing substantial variation in post lengths.

We can see that the SCT annotations perform better than the meddra and also outperform the orinigal folder annotations which are the human annotation notes. From the barplot we can also observe that SpanBERT, both normal and finetuned outperform the other models, even the BioBERT and ClinicalBERT

5.2 Error Analysis Findings

True / Pred	O	B-ADR	I-ADR
O	11772	115	473
B-ADR	97	925	55
I-ADR	394	41	1910

Figure 5: Token-Level Confusion Matrix (SCT merge, BERT)

The qualitative error analysis was conducted on a representative document from the SCT (merge) experiment using BERT. my findings are summarized as follows:

Token Boundary Accuracy: The model correctly identifies the majority of tokens within ADR mentions. For example, the tokens "lip" and "##itor" are consistently labeled as B-ADR and I-ADR, respectively, indicating that the model reliably captures the core parts of ADR expressions.

Boundary Misclassifications: Nevertheless, the confusion matrix computed over the test set reveals that errors tend to occur at the boundaries of ADR spans. In several cases, tokens that should be classified as non-ADR (O) are erroneously marked as B-ADR or I-ADR, and vice versa. This suggests that while the merge preprocessing strategy improves overall span coverage, there is residual ambiguity in determining the exact boundaries of complex, discontinuous annotations.

Confusion Patterns: The aggregated confusion matrix shows that most errors occur between non-ADR tokens and ADR boundary tokens. Such misclassifications potentially lower recall and F1 scores. The need for improved token alignment or post-processing methods is evident, as refining the boundary detection mechanism could further enhance overall performance.

Overall Implications: These error analysis findings underscore that, despite the strong performance of state-of-the-art transformer models, accurately delineating the boundaries of multi-span entities remains challenging. Future work may focus on context-aware post-processing and specialized boundary detection techniques to mitigate these errors.

6 Discussion of Results

my experiments reveal several key insights:

- **Annotation Scheme Impact:** The three annotation schemes yield distinct performance profiles. In particular, the SCT scheme, especially when processed with the merge strategy, appears to yield more consistent and coherent label boundaries. This consistency is reflected in slightly higher F1 scores compared to the Original and MedDRA schemes. The improved boundary consistency in SCT may be due to its mapping to formal clinical terminologies (SNOMED CT), which can reduce ambiguity in span definitions. By contrast, while the Original scheme often achieves competitive precision and recall, it suffers from greater variability because it is based solely on human judgment without the benefit of standardized terminology, while MedDRA performs poorly showing the importance of sct and human annotations.
- **Preprocessing Strategy:** Across all annotation schemes, the merge strategy (which consolidates discontinuous spans into one continuous span) generally produces marginal improvements in F1 scores compared to the split strategy. This suggests that merging helps mitigate the fragmentation of ADE mentions—an issue that can occur when multi-span annotations are treated independently. Such merging enhances the model's ability to capture the complete context of the adverse drug reaction, thereby reducing noise in token-level labels.
- **Model Comparison:** Domain-specific models such as BioBERT and ClinicalBERT consistently outperform the base BERT model across both the MedDRA and SCT schemes, confirming that pretraining on biomedical or clinical corpora yields representations that are better suited for ADE extraction. Notably, the SpanBERT variants stand out: even the un-finetuned (baseline) version of SpanBERT demonstrates

competitive performance, sometimes even rivaling or exceeding that of the finetuned version. This finding suggests that the architecture of SpanBERT, which is designed to enhance span-level representations, is inherently effective for tasks requiring precise span detection. It also implies that with minimal additional task-specific fine-tuning, SpanBERT may offer an efficient solution for ADE extraction.

- **Error Analysis:** my token-level error analysis indicates that, while the models are generally successful in identifying core ADR tokens, the major errors occur at the boundaries of these spans. The confusion matrix reveals that non-ADR tokens are sometimes mislabeled as part of an ADR and vice versa. These boundary misclassifications can result in subtle decreases in recall and F1 scores. This behavior underscores a persistent challenge in token-level annotation alignment and suggests that further refinement—such as context-aware post-processing or improved boundary detection methods—could further boost performance.

7 Conclusion

This work set out to address several research questions critical to improving ADE extraction from noisy, consumer-generated text:

1. **RQ1:** *How do transformer models pretrained on general-domain versus biomedical/clinical corpora perform for ADE extraction?*
   The results indicate that domain-specific models (BioBERT, ClinicalBERT) significantly outperform the base BERT model. This confirms that pretraining on biomedical and clinical texts equips models with a richer understanding of specialized terminology, which is essential for handling noisy input.

2. **RQ2:** *How does the choice of annotation scheme affect the ability of models to capture complete ADE spans?*
   Comparative analyses reveal that the SCT annotation scheme, especially when paired with a merge strategy, provides more consistent label boundaries and yields slightly higher overall F1 scores. The standardized clinical mappings in SCT reduce ambiguity, making it a preferable option for ADE extraction tasks.

3. **RQ3:** *What is the impact of different preprocessing strategies on token-label alignment and downstream performance?*
   The merge strategy consistently improves token-label alignment and performance by consolidating discontinuous annotations, thereby capturing the full context of ADE mentions. This leads to improvements in precision and F1 scores compared to the split strategy.

4. **RQ4:** *Can SpanBERT, even in its baseline form, provide a robust solution for span-level ADE extraction?*
   Results demonstrate that even the un-finetuned SpanBERT performs competitively, at times rivaling its finetuned counterpart and also outperforming the domain-specific models which could show that SPANbert performs really well on ADE extraction and could be useful for future tasks, which also indicates that SpanBERT's architecture, which emphasizes span-level representations, is inherently well-suited for tasks that demand precise boundary detection.

Overall, my study demonstrates that the choice of annotation scheme and preprocessing strategy are critical factors in the performance of transformer-based models for ADE extraction. The superior performance of SpanBERT and the promising results from even its baseline performance—underscore the importance of specialized architectures and pretraining for span-level tasks. Nevertheless, the persistent challenge of accurately delineating entity boundaries, as evidenced by the error analysis, suggests that further work in context-aware post-processing and boundary refinement is warranted. These improvements could enhance the reliability of ADE extraction systems and ultimately contribute to more effective pharmacovigilance in clinical settings.

7.1 Future Work

Future research should focus on improving boundary detection by integrating context-aware post-processing techniques and exploring attention-based models specifically optimized for multi-span entities. Additionally, ensemble methods that combine predictions from diverse domain-specific models (e.g., BioBERT, ClinicalBERT, and SpanBERT) may further boost performance. Investigating minimal fine-tuning strategies for SpanBERT, given its strong baseline performance, and incorporating external clinical knowledge through ontologies could also enhance ADE extraction in noisy texts.

8 Contributions

The paper was entirely written by me (Monish Shah) as I was not able to contribute with my teammate due to health issues and had to shift to a new group.

References

[1] Karimi, S., Metke-Jimenez, A., Kemp, M., & Wang, C. (2015). CADEC: A Corpus of Adverse Drug Event Annotations in Consumer Reviews. *Journal of Biomedical Informatics, 58*, 31–42. https://doi.org/10.1016/j.jbi.2015.03.010

[2] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT*.

[3] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics, 36*(4), 1234–1240.

[4] Alsentzer, E., Murphy, J. R., Boag, W., Weng, W.-H., Jin, D., Naumann, T., & McDermott, M. B. (2019). Publicly Available Clinical BERT Embeddings. *arXiv preprint arXiv:1904.03323*.

[5] Joshi, M., Levy, O., Zettlemoyer, L., & Weld, D. (2020). SpanBERT: Improving Pre-training by Representing and Predicting Spans. In *NAACL-HLT*.

[6] Snow, R., O'Connor, B., Jurafsky, D., & Ng, A. Y. (2008). Cheap and fast – but is it good? Evaluating non-expert annotations for natural language tasks. In *EMNLP*.

[7] Scaboro, S., Portelli, B., Chersoni, E., Santus, E., & Serra, G. (2023). Extensive Evaluation of Transformer-based Architectures for Adverse Drug Events Extraction. *arXiv preprint arXiv:2306.05276*.

[8] Dong, X., Guo, Y., Liu, Q., Patterson, B., & Hong, J. (2024). BERT-based Language Model for Accurate Drug Adverse Event Extraction from Social Media Data. *Frontiers in Public Health*. Retrieved from https://www.frontiersin.org/journals/public-health/articles/10.3389/fpubh.2024.1392180/full.

[9] [Anonymous]. (2022). Adverse Drug Event Detection Using Natural Language Processing. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9810201/.

[10] [Anonymous]. (2021). BERT Prescriptions to Avoid Unwanted Headaches. Retrieved from https://aclanthology.org/2021.eacl-main.149/.

[11] [Anonymous]. (2018). Annotation and Detection of Drug Effects in Text for Pharmacovigilance. *Journal of Cheminformatics*. Retrieved from https://jcheminf.biomedcentral.com/articles/10.1186/s13321-018-0290-y.