Amali Jayatileke | Vidal Mendoza Tinoco | Kelsey Strachan

Article Comparison: An Analysis of News Deja Vu Performance

1. Abstract

The performance of News Deja Vu was evaluated, a semantic search tool that employs named entity recognition (NER) to uncover parallels between present day news stories and historical newspaper articles. We replicated the original pipeline with a Custom Historical NER model trained on texts (1922–1977), using RoBERTa-Large and Hyperband for hyper-parameter optimization. Building on this baseline, we ran four experiments: 1) corpus size validation on progressively smaller subsets 2) comparison against public NER benchmarks: BERT-Large, DistilBERT and RoBERTa-Large 3) topic level evaluation for the ten categories with the highest F1 4) and a hyper-parameter sensitivity study (batch size, learning rate and epochs) on a reduced 2,000 sample corpus.

The historical model consistently delivered the highest precision (0.81) across all scenarios, but at the cost of coverage (recall ~0.04). By contrast, BERT-Large and DistilBERT achieved overall F1 scores of 0.43 and 0.41, respectively. Fine-tuning revealed that small batches, learning rates of 3e-5 and 5e-5 and up to five epochs systematically increased F1 to 0.90, with gains plateauing beyond ten epochs. Topic based tests exposed performance biases, the historical model led in precision for Arts and Entertainment, Community Development and Housing, and Science and Technology, yet it still lagged in recall.

These findings reinforce the benefits of domain specific pre-training and the necessity of rigorous hyper parameter tuning for NER on OCR-noisy historical texts. Future work should probe the model's generalisation on contemporary benchmarks and multilingual historical corpora, and expand the annotated dataset to boost recall without sacrificing precision.

2. Introduction

The News Deja Vu model integrates Named Entity Recognition (NER) within a Natural Language Processing (NLP) framework to identify and classify named entities within text. In parallel, it employs a bi-encoder architecture for semantic similarity analysis, wherein two distinct text inputs are independently transformed into high-dimensional vector representations. These vectors are then compared to assess the degree of semantic similarity between the inputs.[1]

The significance of this work lies in its potential to contribute to the evaluation and understanding of cultural and linguistic evolution. By enabling analysis of how the connotation, frequency, and contextual use of specific terms shift over time – a phenomenon known as semantic drift – this approach can offer meaningful insights into the dynamic nature of language. Moreover, this model facilitates longitudinal comparison of semantic fields across different time periods, which may reveal changes in societal values, moral frameworks, or cultural priorities.

Beyond linguistic insights, the model also holds potential promise for applications in economic and geopolitical analysis. It could help identify patterns in narrative framing and language usage that precede significant historical events such as economic recessions or geopolitical conflicts. This capability could support predictive analytics by uncovering recurring linguistic indicators or shifts in discourse.

Additionally, this model could allow for the tracking of narrative evolution in public discourse, such as media coverage of major events. It could also be leveraged to detect rhetorical strategies associated with emerging patterns of bias, contributing to broader efforts in media literacy and disinformation detection. Finally, its application to legal and policy documents could reveal how legislative language and policy framing evolves over time, offering valuable perspectives on regulatory trends and governance priorities.

The researchers' approach ensured that semantic similarity comparisons between modern and historical articles are not skewed by surface-level name matches. This allows the system to identify conceptually similar stories across time, independent of specific people, organizations, or locations. However, applying NER to historical texts also presents some unique challenges. These texts often feature archaic language along with OCR (Optical Character Recognition) errors introduced during the digitization of scanned documents. These factors complicate accurate entity recognition, underscoring the need for specialized NER models tailored to the linguistic and technical characteristics of historical archives.

3. Methodology

The model developed in this study uses deep learning methods to process historical newspaper data and leverages a custom-trained Sentence-BERT (S-BERT) MPNET model. To emphasize thematic content over specific references, it includes functionality to identify and mask named entities (Figure 1). Both the dataset and the pre-trained model code were accessed through Dr. Melissa Dell's publicly available GitHub repository, Newswire.[2]

The researchers fine-tuned a RoBERTa-Large model, which is an optimized variant of BERT that improves performance by training for longer, with larger batch sizes, on more data, and through removal of the Next Sentence Prediction objective.[3] Randomly selected articles from off-copyright newspapers published between 1922 and 1977 were used as training data. To optimize hyperparameters, they utilized the Hyperband optimization algorithm, which accelerates random search by adaptively allocating resources and early-stopping underperforming configurations.[4] The final model was trained with a learning rate of 4.7e-5, a batch size of 128, over 184 epochs.

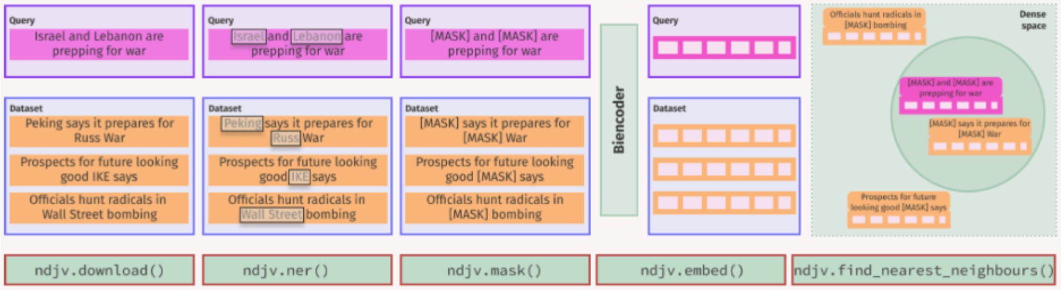
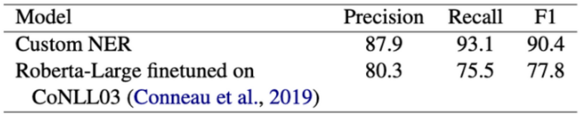

Query
Israel and Lebanon are prepping for war
Dataset
Peking says it prepares for Russ War
Prospects for future looking good IKE says
Officials hunt radicals in Wall Street bombing
Query
Israel and Lebanon are prepping for war
Dataset
Peking says it prepares for Russ War
Prospects for future looking good IKE says
Officials hunt radicals in Wall Street bombing
Query
[MASK] and [MASK] are prepping for war
Dataset
[MASK] says it prepares for [MASK] War
Prospects for future looking good [MASK] says
Officials hunt radicals in [MASK] bombing
Biencoder
Query
Dataset
Officials hunt radicals in [MASK] bombing
Dense space
[MASK] and [MASK] are prepping for war
[MASK] says it prepares for [MASK] War
Prospects for future looking good [MASK] says
ndjv.download()
ndjv.ner()
ndjv.mask()
ndjv.embed()
ndjv.find_nearest_neighbours()

Figure 1. Model Architecture[1]



Model	Precision	Recall	F1
Custom NER	87.9	93.1	90.4
Roberta-Large finetuned on CoNLL03 (Conneau et al., 2019)	80.3	75.5	77.8

Figure 2. Researcher Outcome from Evaluation of NER models[1]

Our project goals are as follows:

1. Performance evaluation of the Custom Historical NER model
2. Comparison of custom model results against publicly available NER models on a shared dataset
3. Comparison of all selected model performance when isolating on specific topics
4. Retraining the model epochs, batch size, and learning rate to attempt to improve performance

The researchers originally obtained 87.9% precision, 93.1% recall, and 90.4% F1 (Figure 2). Our project limitations included the fact that the dataset is tailored to the Custom Historical NER model, which reduced generalizability. Furthermore, the NER model was trained on historical news and, therefore, may reflect biases from a specific period of time. Lastly, the custom bi-encoder architecture is proprietary and so was not publicly available thereby limiting the scope of our project. The alternative models chosen for assessment were the RoBERTa-Large model (2019), the BERT Large (2023) and the DistilBERT (2023). The RoBERTa-Large is an optimized method for pre-training BERT models for performance improvement.[3] RoBERTa-Large hyperparameters are as follows: 24 layers; 1024 hidden; 16 attention heads; ~355M parameters.[5] BERT Large follows the original BERT architecture, which introduces bidirectional training of transformers using a masked language modeling (MLM) objective and next sentence prediction. The hyperparameters for this model are 24 layers; 1024 hidden; 16 attention heads; ~340M parameters.[5] Lastly, the DistilBERT model is a smaller, faster, and lighter version of BERT developed using knowledge distillation. It retains 97% of BERT's performance while reducing the number of parameters. The hyperparameters are 6 layers; 768 hidden; 12 attention heads; ~65M parameters.[5]

4. Experiments

4.1. Exploratory Data Analysis

Through an initial exploratory data analysis, we found that the custom dataset was packed with articles addressing *International Affairs*, *Defense*, and *Law, Crime and Family Issues* (Figure 3). Additionally, it was found that Entity type `O` made up most of the topics ("O": Outside of any named entity - this word is not part of any recognized entity) (Figure 3).

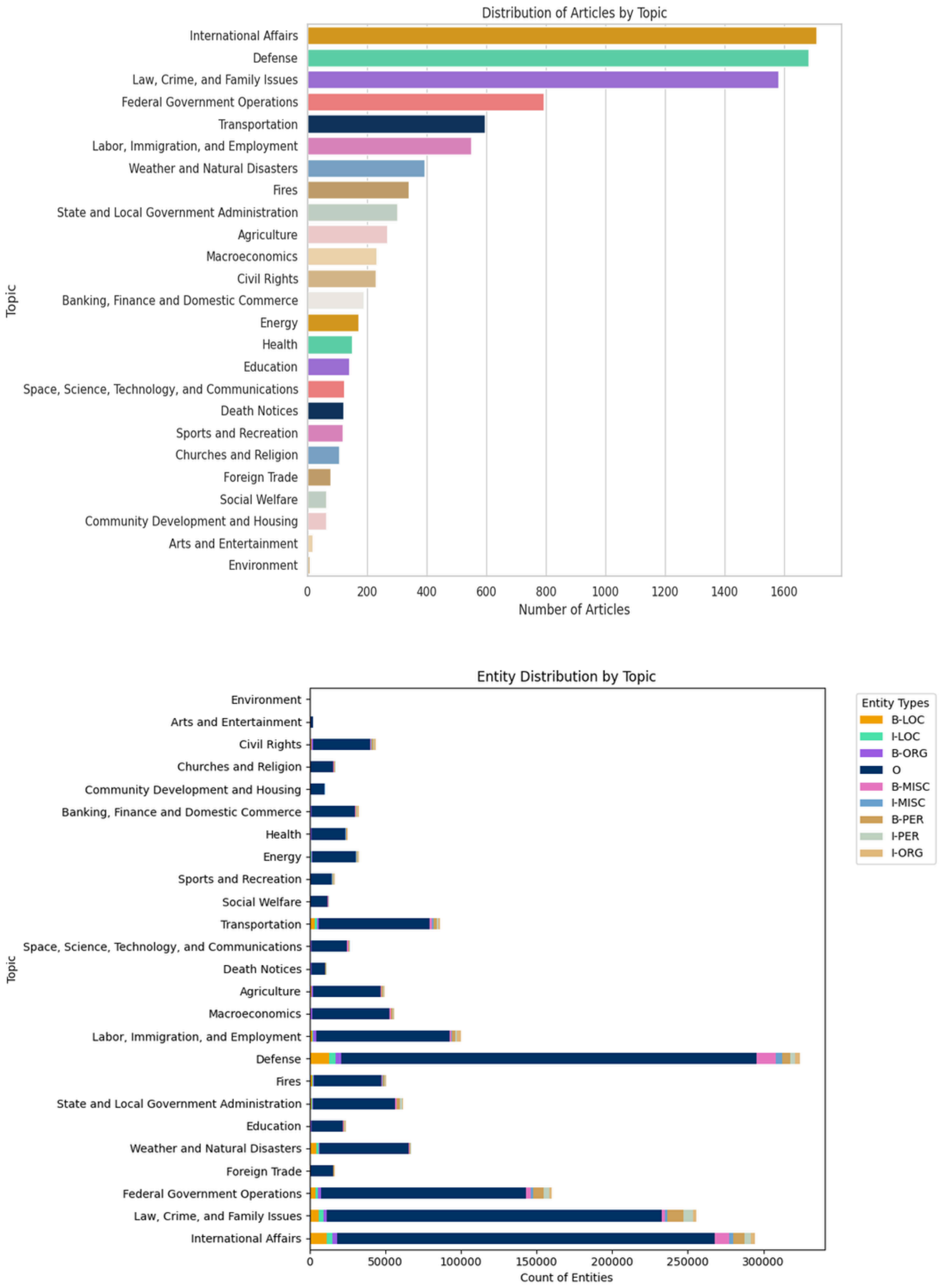


Figure 3. Distributions of Articles (top panel) and Entity Distribution (bottom panel) by Topic

4.2. Experiment 1: Assessing Model Performance on Reduced Corpus

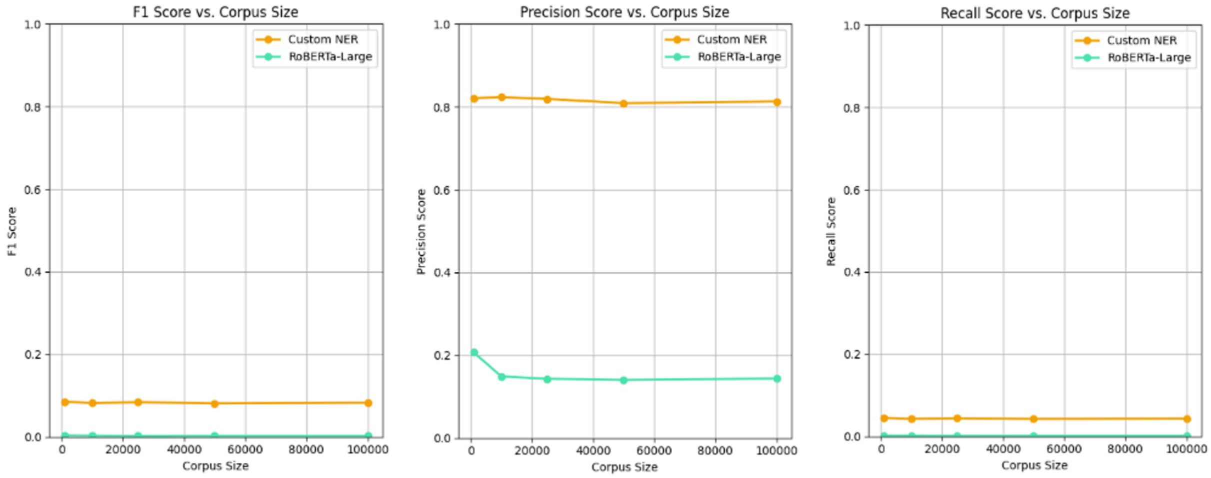


Figure 4. Reproduction of Validation Results

Models were evaluated using a selection of corpus sizes (1,000, 10,000, 25,000, 50,000, 100,000). The Custom Historical NER model consistently outperformed the RoBERTa-Large model across all metrics (F1 score, Precision, and Recall) for each corpus size tested. Subject to the corpus sizes tested, the Custom Historical NER model achieved Precision values close to those reported in the research article (~80%). However, other metrics for both the Custom Historical NER and RoBERTa-Large models were comparatively lower than those reported in the research article (Figure 2).

4.3. Experiment 2: Assessing Performance Against Alternative Models

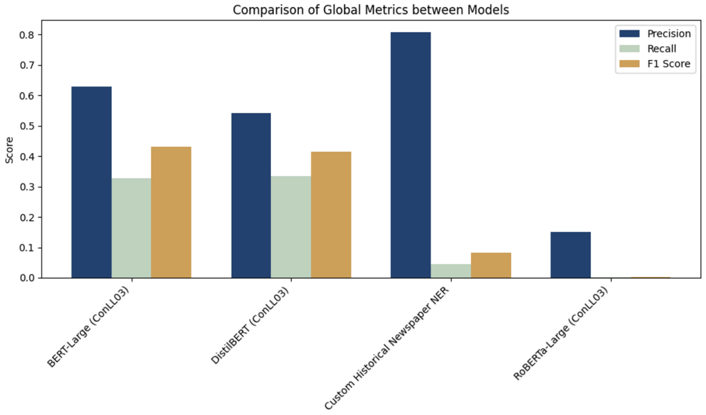


Figure 5. Comparison of Global Metrics between Models

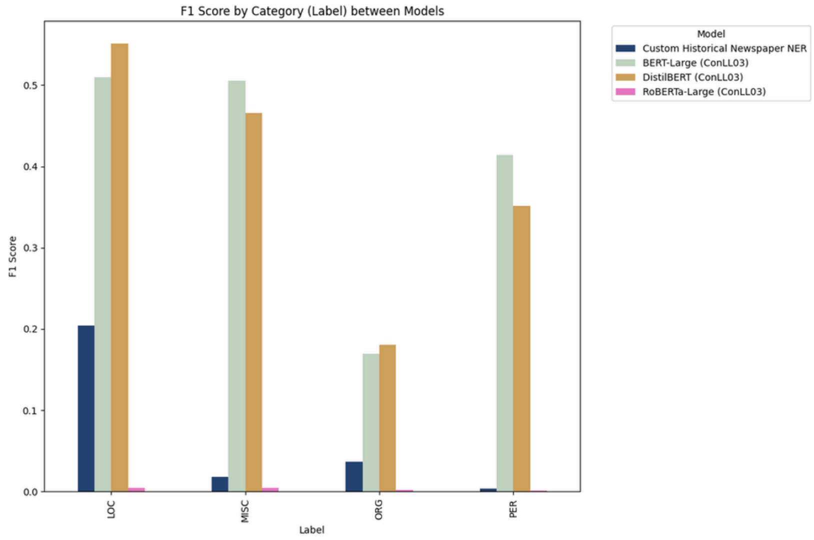


Figure 6. F1 Score by Category between Models

Four NER models were evaluated on the digitized historical corpus, with the aim of comparing their performance in terms of precision, recall, and F1 score. The BERT-Large (CoNLL03) model achieved the best overall performance with an F1 score of 0.43, closely followed by DistilBERT (CoNLL03) with an F1 score of 0.41. Both models achieved a reasonable balance between precision and coverage, making them suitable for NER tasks in partially noisy contexts.

The Custom Historical NER model, specifically designed for historical texts with OCR noise, achieved very high precision (0.81) but with extremely low recall (0.04), resulting in an F1 score of ~0.08. This behavior indicates that the model is highly conservative, identifying entities only when it is very confident, which drastically reduces its practical usefulness. For its part, the RoBERTa-Large (CoNLL03) model practically failed to identify entities in this corpus, with an F1 of only 0.003, suggesting a significant lack of adaptation to the historical domain.

When breaking down the results by entity type, we see that DistilBERT performs best in detecting locations (LOC) (F1 = 0.55) and organizations (ORG) (F1 = 0.18), even outperforming BERT-Large in these categories. In contrast, BERT-Large maintains the best ability to identify people (PER), with an F1 of 0.41. Both modern models also show strong results in the miscellaneous category (MISC).

The custom model, although trained to work with ancient texts, performed very poorly in all categories. Its best result was for locations, with an F1 of just 0.20, and its worst performance was for people (F1 = 0.0038). This reinforces the concept that, while its accuracy is high, its coverage is insufficient for practical NER tasks.

4.4. Experiment 3: Assessing Model Performance on Data Subsection

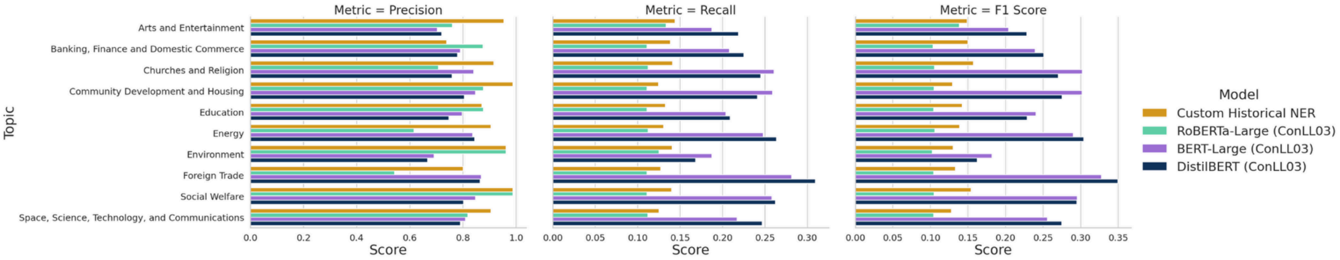


Figure 7. Comparison of Metrics between Models based on Topic

When we selected data points representing the top 10 highest metric values across all models and topics, we found that within this subset, the Custom Historical NER showed improved performance over RoBERTa-Large for F1-Score, Precision, and Recall. However, it still underperformed in comparison to BERT-Large and DistilBERT. The Custom Historical NER was found to have the highest precision amongst all models across the top 10 topics isolated (highest precision: *Arts and Entertainment*, *Community Development and Housing*, *Churches and Religion*, and *Space, Science, Technology and Communications*). Overall, the transformer-based models, particularly DistilBERT and RoBERTa-Large, demonstrated improved performance compared to the Custom Historical NER model, as indicated by their higher F1-Scores across most topics. While Precision scores tend to be higher than Recall scores for all models, implying a tendency to make fewer false positive predictions than missing actual entities, the F1-Score highlights DistilBERT as often achieving the best overall results. Notably, the performance of each model varied across different topics. For example, DistilBERT excelled in identifying entities within the *Foreign Trade* domain, while RoBERTa-Large appears to perform strongly in *Science, Technology, and Communications.* Conversely, the Custom

Historical NER model appears to struggle more significantly with topics like *Community Development and Housing.*

4.5. Experiment 4: Retraining Model and Assessing Performance Against Alternative Models

In our final experiment, we explored the sensitivity of model performance to key hyperparameters — specifically batch size, learning rate, and number of training epochs. While acknowledging that the Custom Historical NER model was originally trained with a learning rate of 4.7e-5, a batch size of 128, over 184 epochs, and that the benchmark NER models used for comparison contain 60 to 350 million parameters, it's important to note that we did not have access to equivalent computational resources. Consequently, we focused our experiment on a reduced corpus of 2,000 samples, compared to the original 2.7 million-row dataset, to stay within practical resource limits.

We evaluated model accuracy using the F1-score, as it provides a balanced measure of both precision and recall. This experiment was designed to help us assess how rapidly performance gains can be achieved through hyperparameter tuning, and to compare the behavior of the Custom Historical NER model against three well-established benchmark models: RoBERTa-Large, BERT-Large, and DistilBERT.

4.4.1. Batch size

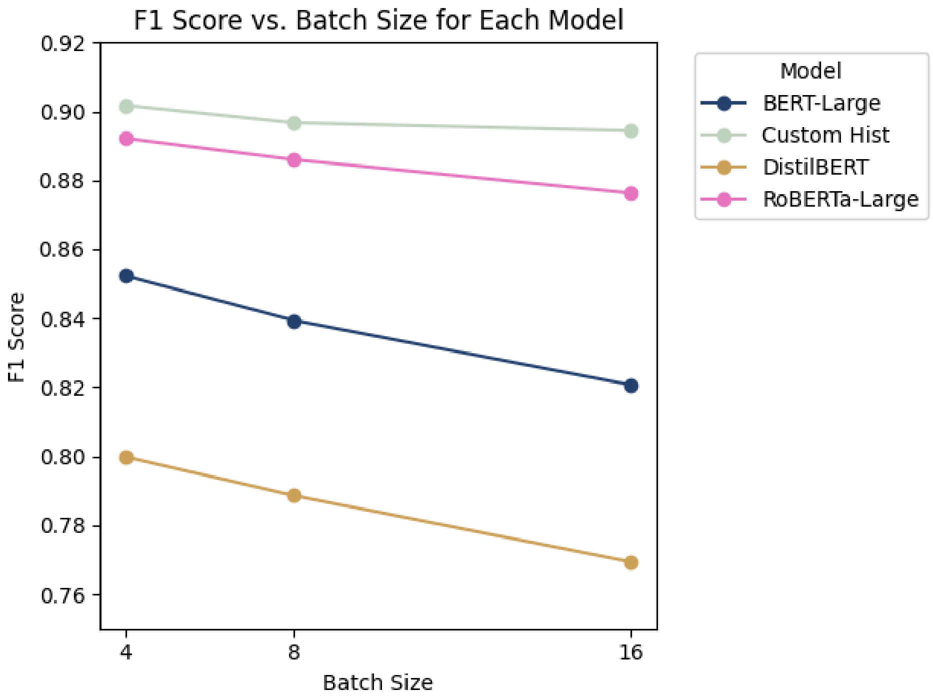


Figure 8. Tuning Batch Size of Custom Historical NER

When analyzing the F1 vs. batch size curve, we observe a clear and consistent pattern across all models: the lower the batch size, the higher the F1. Batch=4 achieves the best results, and as we increase to 8 and 16, performance gradually declines. This suggests that the noisy gradients of smaller batch sizes help the model generalize better in this NER corpus.

The impact is more pronounced in the lighter models. DistilBERT sees a drop of almost 3 F1 points between batch=4 (0.800) and batch=8 (0.789), and another 1.9 points when moving to 16 (0.769). In contrast, the very robust Custom Historical NER only loses 0.7 points when moving from batch=4 (0.902) to batch=16 (0.894), showing lower sensitivity. For the two larger weights, the decline is intermediate. RoBERTa-Large drops from 0.892 (batch=4) to 0.876 (batch=16), and BERT-Large from 0.852 to 0.821.

4.4.2. Learning Rate

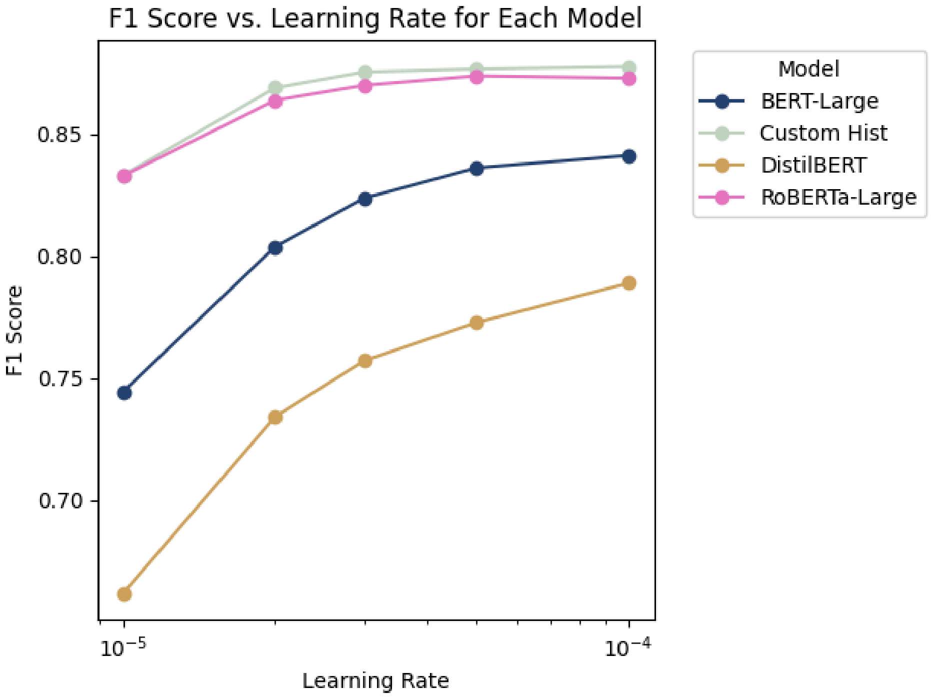


Figure 9. Tuning Learning Rate of Custom Historical NER

At the lowest learning rate of 1e-5, BERT-Large achieves an F1 score of ~0.74, the Custom Historical NER and RoBERTa-Large are similar starting around ~0.83, and we see that DistilBERT outputs the lowest value at ~0.66 (Figure 9). As the learning rate increases to 3e-5, all models improve: BERT-Large (0.80), the Custom Historical NER (0.87), DistilBERT (0.73), and RoBERTa-Large (0.86). Further increasing the learning rate to 5e-5 sees BERT-Large at ~0.82, Custom Historical NER increasing to 0.88, DistilBERT reaching ~0.76, and RoBERTa-Large achieving about 0.87. Finally, at the highest tested learning rate of 1e-4, BERT-Large plateaus around 0.84, the Custom Historical NER remains high at ~0.88, DistilBERT reaches its peak of ~0.79, and RoBERTa-Large also plateaus around ~0.88. We see that while all models benefit from increasing the learning rate initially, the optimal rate varies. We find that the Custom Historical NER model consistently achieves the highest F1 scores, reaching ~0.88 at learning rates of 5e-5 and 1e-4. In contrast, DistilBERT consistently underperforms, with a peak F1 score of ~0.79. BERT-Large and RoBERTa-Large demonstrate competitive performance in the higher learning rate ranges, achieving F1 scores around 0.84 and 0.88 respectively at 1e-4.

4.4.3. Epochs

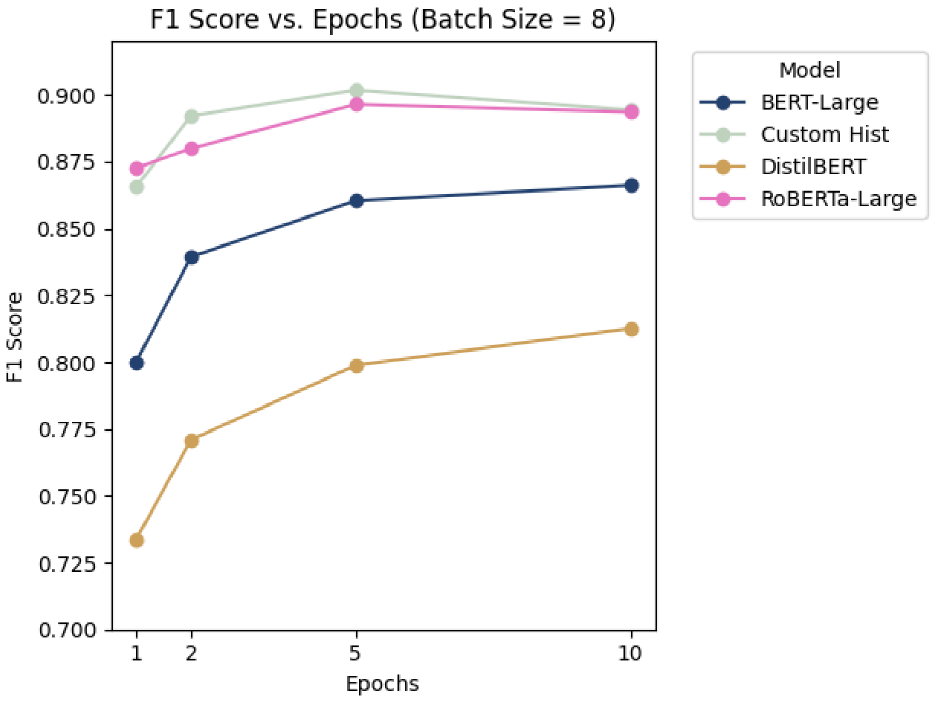


Figure 10. Tuning Epochs of Custom Historical NER

Figure 10 illustrates the relationship between the number of training epochs and the resulting F1-scores for several models fine-tuned on the custom historical dataset. At one epoch, BERT-Large achieves an F1-score of approximately 0.80, while the Custom Historical NER model starts around 0.86. RoBERTa-Large begins slightly higher at 0.87, and DistilBERT notably lags behind at around 0.73. As training progresses to two epochs, performance improves across all models. BERT-Large rises to 0.84, Custom Historical NER jumps to 0.89, DistilBERT increases to 0.77, and RoBERTa-Large reaches 0.89, though it still marginally trails the Custom Historical NER model at this stage. By five epochs, BERT-Large edges up to an F1-score of approximately 0.86, while Custom Historical NER makes a marginal gain to nearly 0.90. DistilBERT continues its steady climb to 0.79, and RoBERTa-Large closely follows at 0.89. At the highest tested setting of ten epochs, both Custom Historical NER and RoBERTa-Large plateau at around 0.89, indicating a convergence in their performance after initial differences. BERT-Large stabilizes at 0.86, and DistilBERT reaches its peak at 0.81, maintaining a noticeable performance gap relative to its larger counterparts.

A few patterns become clear from these results. The Custom Historical NER model consistently outperformed the other models through epochs two and five, achieving the highest F1-scores in the earlier stages of training. However, by ten epochs, RoBERTa-Large closed the gap, achieving performance levels on par with the Custom Historical NER model. While DistilBERT showed consistent improvement over epochs, its performance remained substantially lower throughout, peaking below 0.82. BERT-Large demonstrated reliable incremental gains but plateaued lower than the top-performing models.

Another key observation is the rate of performance improvement varied across model architectures, with diminishing returns observed for most models after five epochs. These outcomes are also sensitive to the random seed used when subsetting rows from the original training data, which can influence results on smaller corpora. Since Custom Historical NER and RoBERTa-Large achieved comparable results by the tenth epoch, it would be worthwhile in future work to explore higher epoch counts, alternative learning rate schedules, or dynamic batch sizing to investigate potential for further gains or to monitor signs of overfitting.

5. Conclusion

Through our project, we successfully reproduced the validation results from the original study and demonstrated that the Custom Historical NER model outperforms RoBERTa-Large, even when trained on a relatively limited historical corpus (Experiment 1). Our experiments highlighted that different models exhibit varying strengths depending on the entity type, reinforcing the need to carefully balance precision and recall for optimal overall performance. The results from Experiment 2 confirmed that combining modern language models with domain-specific pretraining on historical texts yields the most effective strategy for historical NER tasks. Additionally, we observed that models pre-trained on modern corpora, such as BERT-Large and DistilBERT, remain surprisingly robust when applied to noisy, OCR-degraded historical texts. However, these models exhibited notable limitations, particularly in identifying infrequent entities or those with altered formats due to OCR errors.

Experiment 3 revealed performance biases within the Custom Historical NER model across different news topics. The highest precision was observed for categories such as *Arts and Entertainment*, *Community Development and Housing*, *Churches and Religion*, and *Space, Science, Technology and Communications*. These variations likely reflect both the model's domain-specific training and inherent topical biases present in news content from 1922–1977.

Findings from Experiment 4 demonstrated that while all models benefit from additional training epochs, increased learning rates, and smaller batch sizes initially, the rate of improvement and eventual performance plateau vary across model architectures. Notably, DistilBERT consistently underperformed compared to its larger counterparts, while BERT-Large and RoBERTa-Large exhibited competitive performance when hyperparameters were optimized. The consistent superiority of the Custom Historical NER model underscores the value of domain-specific pretraining and highlights the critical importance of rigorous hyperparameter tuning in specialized tasks such as historical entity recognition.

For future work, we recommend evaluating the Custom Historical NER model on established contemporary NER benchmarks such as CoNLL-2003 (Hugging Face)[6], which includes widely used entity classes like *PER*, *LOC*, *ORG*, and *MISC*. This would clarify whether the model’s performance gains are limited to historical or OCR-degraded contexts or if they generalize to contemporary NER tasks. Additionally, assessing the model on specialized historical/OCR datasets such as HIPE-2022 (Hugging Face)[6] would provide further insight into its robustness across diverse, noisy archival corpora. Another valuable direction would be to evaluate the model's adaptability and cross-lingual transfer potential on non-English historical corpora. Although this was beyond the scope of the current project due to the limited availability of labeled non-English historical NER datasets, it remains an important area of ongoing research.

Finally, to further enhance the model's recall without compromising precision, adjustments to the confidence threshold or retraining with a larger, more diverse annotated corpus are recommended. Collectively, our findings suggest that a hybrid strategy—combining finely tuned modern transformer-based models with domain-specific pretraining on historical data—offers the most promising pathway to improving named entity recognition performance on digitized historical newspaper archives.

References

[1] B. Franklin, E. Silcock, A. Arora, T. Bryan, and M. Dell, "News Deja Vu: Connecting Past and Present with Semantic Search," arXiv preprint arXiv:2406.15593, Jun. 2024. [Online]. Available: https://arxiv.org/abs/2406.15593arXiv

[2] Dell-Research-Harvard, Newswire: A Large-Scale Structured Database of a Century of Historical News, GitHub repository, [Online].
Available: https://github.com/dell-research-harvard/newswire. [Accessed: Apr. 5, 2025].

[3] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," arXiv preprint, arXiv:1907.11692, Jul. 2019. [Online]. Available: https://arxiv.org/abs/1907.11692. [Accessed: Apr. 19, 2025].

[4] L. Li, K. Jamieson, G. DeSalvo, A. Rostamizadeh, and A. Talwalkar, "Hyperband: A Novel Bandit-Based Approach to Hyperparameter Optimization," arXiv preprint, arXiv:1603.06560, Mar. 2016. [Online]. Available: https://arxiv.org/abs/1603.06560. [Accessed: Apr. 19, 2025].

[5] T. Wolf et al., "Transformers: State-of-the-Art Natural Language Processing," Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Online, Oct. 2020, pp. 38–45. [Online]. Available: https://www.aclweb.org/anthology/2020.emnlp-demos.6

[6] Q. Lhoest et al., "Datasets: A community library for natural language processing," in *Proc. 2021 Conf. Empirical Methods in Natural Language Processing: System Demonstrations*, Online and Punta Cana, Dominican Republic, Nov. 2021, pp. 175–184. [Online]. Available: https://aclanthology.org/2021.emnlp-demo.21