Improving Named Entity Recognition in Song Lyrics Through Domain-Adaptive Pretraining and Genre-Specific Fine-Tuning

Natural Language Processing and Deep Learning - BSNLPDL2KU
IT University of Copenhagen

Aniela Marta Ciecierska anci@itu.dk
Jakub Piotr Gasior jaga@itu.dk
Maja Rzeszotarska marz@itu.dk

Abstract

Creating a Named Entity Recognition (NER) model that performs well on domain-specific datasets can be very challenging, especially since most NER models are trained on formal or general-purpose data without being tailored to a particular domain. In this project, we focus on a dataset of song lyrics. Song lyrics differ significantly in language style, often including creative spelling, slang, abbreviations, and pop culture references—posing a considerable challenge for entity recognition tasks.

In this paper, we explore how to improve the performance of an NER model originally trained on general-purpose data by applying domain-adaptive pretraining, fine-tuning, and continuous learning using lyrics from three music genres: pop, country, and rap/hip-hop. We compare several models and evaluate their contribution to performance improvements. Our results show that domain adaptation significantly outperforms the baseline general-purpose model, achieving an improvement of 7.19 percentage points in span F1 score. These findings highlight the importance of domain adaptation for robust entity recognition in creative and informal datasets. [1]

1 Introduction

In Named Entity Recognition (NER) models, the accurate identification of named entities is crucial. For optimal performance, these models must effectively capture the unique characteristics of the data they are trained on. When data, such as song lyrics, contain distinct features, such as informal language, slang, references to well-known figures and brands, and poetic or abstract expressions, traditional NER models trained on general-purpose datasets may struggle to capture these nuances. As a result, their performance may lack robustness and accuracy.

Techniques such as domain-adaptive pretraining and fine-tuning have been proven to improve the performance of NER models, but how effective are these techniques when applied to song lyrics? Specifically, to what extent does domain-adaptive pretraining and fine-tuning on song lyrics enhance NER model performance compared to general-purpose model, and do certain music genres contribute more effectively to this adaptation?

2 Related Work

Prior work has demonstrated the effectiveness of domain-adaptive pretraining (DAPT) and task-adaptive pretraining (TAPT) in improving performance on downstream NLP tasks. Gururangan et al. (2020) show that continued pretraining in unlabeled domain-specific data significantly boosts performance, even when using large-scale models such as RoBERTa. Sachidananda et al. (2021) further highlight the importance of adapting tokenization and pretraining to domain-specific lexical properties, especially in multilingual contexts. Although these studies target formal or technical domains, such as biomedical or financial texts, our work focuses on a significantly underexplored genre: song lyrics. This genre poses unique challenges for named entity recognition (NER) due to its creative use of language, informal tone, and frequent use of slang and cultural references. Although prior research Song and Beck (2023) has explored the modeling of emotions in lyrics, they do not involve structured NER tasks or manually annotated data. Similarly, Hong et al. (2023) explore domain relevance via masked domain-specific keywords, yet still operate on

[1] https://github.com/aniela2906/NLP-spring-2025.git

large, structured corpora and do not address annotation or evaluation in informal domains. In contrast, we manually annotate a novel dataset of song lyrics with standard NER tags, and systematically evaluate both general-purpose and lyrics-specific models. To our knowledge, this is the first work that directly compares general and domain-specific NER models on a manually annotated song lyrics.

3 Data

To construct our datasets, we first identified popular artists in three music genres—pop, country, and rap/hip-hop—using rankings from Kworb, Vault, and Browne et al. (2017). We then selected 50 artists per genre, and for each artist, we chose 20 songs based on popularity from Kworb and lyric availability on Genius (b). We scraped the lyrics for all selected songs using the Genius API (Genius, a), resulting in 1,000 songs per genre. We preprocessed the lyrics by removing duplicate or repetitive lines commonly found in songs. Finally, the lyrics were tokenized and converted into the IOB2 format.

4 Methodology

4.1 Baseline Model

Our baseline model is based on the deepset/roberta-base-squad2 transformer—a RoBERTa model (Liu et al. (2019), Wolf et al. (2020)) originally fine-tuned for question answering. We adapted it for token classification using the Mayhew (2022b)English Web Treebank (EWT) dataset. The model is trained with standard NER supervision, including sentence-level tokenization, BIO tagging, and alignment of labels to word-piece tokens.

To strengthen our in-domain baseline, we applied Domain-Adaptive Pretraining (DAPT), which involved masked language modeling (MLM) on a large corpus of song lyrics prior to fine-tuning the model for NER. All subsequent models (pop, country, rap/hip-hop, and the combined three-genre model) were initialized from this DAPT model, ensuring a consistent starting point and enabling fair comparisons across training strategies.

4.2 Manual Annotations

Since many sentences in our datasets did not contain any named entities, we first used our DAPT (Domain-Adaptive Pretraining) model to automatically label the data and identify sentences likely to include named entities. Based on this initial labeling, we sampled and manually annotated 2,000 sentences from each music genre dataset (pop, country, and rap/hip-hop), following the annotation guidelines provided by Stephen Mayhew Mayhew (2022a). The annotation scheme was consistent with that of the EWT dataset, using the entity tags: B-PER, I-PER, B-ORG, I-ORG, B-LOC, I-LOC, and O. After annotation, the data were shuffled and randomly split into training and test sets. Each training file contains 1,000 sentences per genre, while the test set consists of 3,000 sentences (1,000 from each genre). In addition, we created a combined training file that included all three genres.

4.3 Fine-Tuning

We fine-tuned our DAPT model, initially trained on the EWT dataset, using the manually labeled datasets for each genre (pop, country, rap/hip-hop) as well as the combined dataset. Next, we used the resulting fine-tuned models to pseudo-label the remaining, unlabeled song lyrics—each model labeling data from its corresponding genre. We then conducted an additional round of fine-tuning using both the manually labeled and pseudo-labeled data for each genre and the combined set. At each stage of this pipeline, we evaluated performance on the lyrics test set to monitor potential improvements in the performance.

4.4 Exploratory Data Analysis

To better understand the characteristics of our data, we conducted an exploratory analysis of both general-purpose and genre-specific datasets. As shown in Table 1, the English Web Treebank (EWT) dataset used for baseline training contains over 229,000 tokens, but a relatively low proportion of named entity tokens (5.16%). In contrast, our manually annotated genre-specific datasets—each comprising 1,000 sentences from pop, country, and rap/hip-hop lyrics—exhibit significantly higher entity densities: country leads with 16.9%, followed by

rap/hip-hop (15.3%) and pop (12.4%). This is a result of intentionally sampling sentences that contain named entities for the training sets. The pseudo-labeled datasets contain far more sentences and tokens overall, but have lower entity densities, with rap/hip-hop showing noticeably higher density than pop or country. Finally, our manually annotated lyrics test set, consisting of 3,000 sentences, has a high entity density of 14.83%, making it a robust benchmark for evaluating model performance. Table 2 further shows the distribution of unique named entities across datasets, with rap/hip-hop containing significantly more unique entities than the other genres.

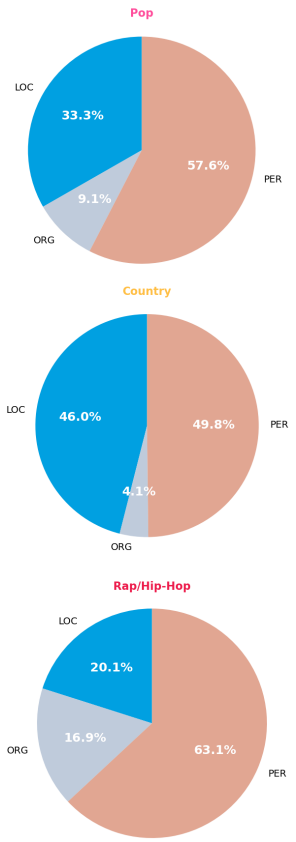

Figure 1: Percentage of each named entity in different genres.

Figure 1 breaks down the distribution of named entity types across genres more. For additional context, a bar chart illustrating the absolute counts of each named entity type can be found in the appendix (see Figure 3). In pop lyrics, person (PER) entities dominate at 57.6%, followed by locations (LOC, 33.3%) and a small part of organizations (ORG, 9.1%). Country lyrics show a more balanced entity profile, with 49.8% PER and a notably high 46.0% LOC entities, reflecting the genre's frequent geographic references. In contrast, rap/hip-hop lyrics are heavily skewed toward PER entities (63.1%), with fewer LOC (20.1%) but a relatively higher proportion of ORG entities (16.9%), possibly reflecting references to brands, labels, or groups. These findings highlight the distinct thematic emphases of each genre and reinforce the importance of genre-specific fine-tuning for effective named entity recognition in lyrical text.

Dataset	PER	ORG	LOC	Unique Named Entity Tokens
EWT data	3,217	1,660	3,111	3390
pop (manually labeled)	510	72	244	499
country (manually labeled)	585	38	528	532
rap/hip-hop (manually labeled)	772	188	235	809
pop (pseudo-labeled)	248	73	59	202
country (pseudo-labeled)	393	64	112	200
rap/hip-hop (pseudo-labeled)	4,511	1,872	1,366	4012
lyrics test	1,807	295	1,066	1704

Table 2: Number of specific named entity tokens in each dataset.

5 Results

To evaluate the performance of different NER models on a dataset with significant class imbalance (the overwhelming majority of "O" tags and unequal distribution of named entity types), relying on accuracy alone would be misleading. Therefore, we use the widely accepted span-level F1 score, which balances precision and recall via the harmonic mean. This metric considers a prediction correct only if both the entity type and the entire span exactly match the gold annotation.

Table 3 presents the evaluation results (bar chart version can be found in the appendix (see Figure 4).The baseline model, trained only on the EWT dataset, achieved a score of 80,28%. The DAPT model, pretrained on the full corpus of lyrics, showed a noticeable improvement with 81.37%. All models fine-tuned on manually annotated song lyrics showed significant gains, with the rap/hip-hop model achieving the highest F1 among the single-genre models (85.19%), followed by pop (85.08%) and country (84.65%). The model trained on all manually labeled lyrics across genres achieved the highest overall F1 score of 87.46%.

Applying continuous learning using pseudo-labeled data led to modest improvements

Dataset	Sentences	Tokens	Named Entity Tokens	Entity-Labeled Token Density (%)	Avg Tokens/Sent.
EWT data	14,544	229,728	11,856	5.16	15.80
pop (manually labeled)	1,000	9,295	1,153	12.4	9.29
country (manually labeled)	1,000	10,167	1,718	16.9	10.17
rap/hip-hop (manually labeled)	1,000	11,439	1,750	15.3	11.44
pop (pseudo-labeled)	33,129	318,566	484	0.15	9.62
country (pseudo-labeled)	18,418	175,577	728	0.41	9.53
rap/hip-hop (pseudo-labeled)	51,141	577,204	11,630	2.01	11.29
lyrics test	3,000	31,094	4,610	14.83	10.36

Table 1: **S**ummary of datasets' characteristics.

across all models, with increases of less than one percentage point (a file containing detailed results for each model is available on our GitHub).

Dataset	Slot-Level F1 Score (%)
EWT (basic)	80.28
EWT (DAPT)	81.37
EWT (DAPT) + pop	85.08
EWT (DAPT) + pop (cont.)	85.51
EWT (DAPT) + country	84.65
EWT (DAPT) + country (cont.)	85.09
EWT (DAPT) + rap/hip-hop	85.19
EWT (DAPT) + rap/hip-hop (cont.)	85.85
EWT (DAPT) + all genres	87.46
EWT (DAPT) + all genres (cont.)	87.47

Table 3: Slot-Level F1 score results for all models.

6 Analysis

As domain-specific data was progressively added, the unique named entity coverage in the test data significantly increased (see table 4). The coverage rose from 13.67% in models trained exclusively on the EWT dataset, to 23.06% in pop model, 25.88% country model, and 23.88% in rap/hip-hop model. Further incorporating pseudo-labeled data led to only slight increases in coverage for pop and country, as these datasets contained fewer named entity tokens. However, the rap/hip-hop model saw a substantial increase, with coverage reaching 38.32%.

As expected, the set combining all genres yielded the highest coverage at 41.14%, with its continuous learning version achieving an impressive 53.11%. These results highlight the importance of domain adaptation. Clearly, song lyrics contain many named entities that are uncommon in universal datasets like EWT, which explains the marked improvement in coverage as more domain-specific data was added.

It is essential to note that while allowing the model to learn specific named entities is beneficial, it is not the same as ensuring that the model can generalize well to unseen data, especially when dealing with differently structured sentences. This distinction is crucial when interpreting the results, as higher coverage does not always correlate with better overall model performance, particularly in terms of generalization to real-world, diverse text.

Dataset	Unique NE Coverage (%)
EWT (basic)	13.67
EWT (DAPT)	13.67
EWT (DAPT) + pop	23.06
EWT (DAPT) + pop (cont.)	24.06
EWT (DAPT) + country	25.88
EWT (DAPT) + country (cont.)	26.88
EWT (DAPT) + rap/hip-hop	23.88
EWT (DAPT) + rap/hip-hop (cont.)	38.32
EWT (DAPT) + all genres	41.14
EWT (DAPT) + all genres (cont.)	53.11

Table 4: Percentage of unique named entities in the test set that are covered by the training data.

To further examine differences in model performance, we analyzed the distribution of error types across models (see Figure 2). We focused on three types of errors: missed entities (when a named entity is not detected), spurious entities (when a non-named entity is incorrectly labeled as such), and wrong labels (when the correct entity is detected but labeled incorrectly). The most noticeable improvement is observed in the missed errors, where the number of missed entities drops significantly after adding domain-specific data (from 791 in the EWT-only model to 210 in the country model).

For both the pop and country models, adding pseudo-labeled data shows a tradeoff: while the number of spurious errors decreases, the number of missed entities increases. Interestingly, this trend is not observed in the rap/hip-hop model or in the all genres combined model, where the errors behave differently. The wrong labels error type seems

to vary the least across models. In most cases, models either fail to detect a named entity or, conversely, are too generous with labeling.

These observations are quite insightful, as they suggest that while continuous learning did not significantly improve performance, it did alter the distribution of errors. A more detailed breakdown of the relative proportions of each error type across models is provided in the appendix (see Figure 5)

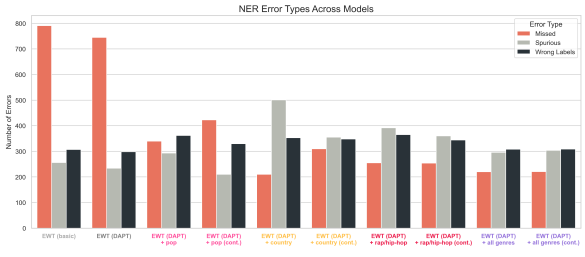

Figure 2: NER error types across models.

7 Discussion

Both domain-adaptive pretraining (DAPT) and fine-tuning on domain-specific data significantly improved model performance, yielding a 7.18 percentage point gain in span-level F1 score. However, adding pseudo-labeled data provided only marginal improvements—less than one percentage point for each model. This suggests that the most impactful factors for improving performance are learning the structure of a domain-specific language and having access to high-quality annotated data.

These findings directly address our research question, demonstrating that domain-adaptive pretraining and genre-specific fine-tuning meaningfully enhance NER performance compared to a general-purpose model. While all three music genres exhibited similar performance trends, with rap/hip-hop achieving slightly higher scores—both with manually labeled and pseudo-labeled data—these differences were small and may be attributable to random variation.

It is important to note that the three genres differ substantially in terms of sentence structure and the density and diversity of named entities. While we specifically sampled sentences containing named entities for annotation, raw unfiltered lyrics reveal that rap/hip-hop typically contains more, and more diverse, named entities than the other genres. This makes rap/hip-hop a particularly rich source for NER tasks. Moreover, rap/hip-hop songs tend to be longer and contain more sentences per track compared to pop or country, making it easier to collect a large and diverse dataset from fewer sources.

Another important factor is the manual annotation process itself. Labeling was performed entirely by us, which is both time-consuming and susceptible to human error. This inherently limits the size of the datasets we could use for training and testing. Additionally, ambiguity in the lyrics presented challenges during annotation—especially in rap/hip-hop, where nicknames, slang, and non-standard language forms are prevalent. While rap/hip-hop offers a wealth of entities, accurately labeling them often requires additional effort and domain knowledge.

8 Conclusion

This study highlights the importance of domain-specific adaptation in natural language processing tasks, particularly for Named Entity Recognition (NER) in the context of song lyrics. Our work shows that domain-adaptive pretraining (DAPT) followed by genre-specific fine-tuning significantly improves model performance, making a span-level F1 score increase of up to 7.19 percentage points compared to the baseline model trained only on general-purpose data. The model trained on all three music genres combined performed best, achieving the highest F1 score and the broadest named entity coverage. All genres contributed comparably to performance gains, with rap/hip-hop showing slightly better results, likely due to its dense and diverse use of named entities. These findings prove that focusing on domain-specific NLP is crucial when dealing with informal, creative datasets like song lyrics, and that training models on the right kind of text and labeling data accurately can lead to strong and reliable NER systems that understand the context better.

References

David Browne, Jon Dolan, Jon Freeman, Chris Parton, Stephen L. Betts, Andrew Leahey, Joseph Hudak, Kory Grow, Marissa R. Moss, Maura Johnston, Joe Levy, Will Hermes, David Cantwell, and Jonathan Bernstein. 2017. 100 greatest country artists of all time. https://www.rollingstone.com/music/music-lists/100-greatest-country-artists-of-all-time-195775. Accessed: March 2025.

Genius. a. Genius api documentation. https://docs.genius.com. Accessed: March 2025.

Genius. b. Genius: Song lyrics and knowledge. https://genius.com. Accessed: March 2025.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Pengfei Hong, Rishabh Bhardwaj, Navonil Majumder, Somak Aditya, and Soujanya Poria. 2023. A robust information-masking approach for domain counterfactual generation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3756–3769, Toronto, Canada. Association for Computational Linguistics.

Kworb. Spotify artist rankings. https://kworb.net/spotify/artists.html. Accessed: March 2025.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. In *arXiv preprint arXiv:1907.11692*.

Stephen Mayhew. 2022a. Universal ner, annotation 407 guidelines. In *Universal NER*.

Stephen Mayhew. 2022b. Universal ner project: English web treebank (ewt). In *Universal NER*.

Vin Sachidananda, Jason Kessler, and Yi-An Lai. 2021. Efficient domain adaptation of language models via adaptive tokenization. In *Proceedings of the Second Workshop on Simple and Efficient Natural Language Processing*, pages 155–165, Virtual. Association for Computational Linguistics.

Yingjin Song and Daniel Beck. 2023. Modeling emotion dynamics in song lyrics with state space models. *Transactions of the Association for Computational Linguistics*, 11:157–175.

Music Metrics Vault. Most streamed artists on spotify. https://www.musicmetricsvault.com/most-streamed-artists-spotify. Accessed: March 2025.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Appendix

Group Contributions: All group members contributed equally to the project. The manual annotation task was divided among members, with each person labeling 2,000 sentences from a specific music genre. The group collaborated on all other aspects of the project, distributing tasks evenly and reviewing each other's work to ensure accuracy and quality.

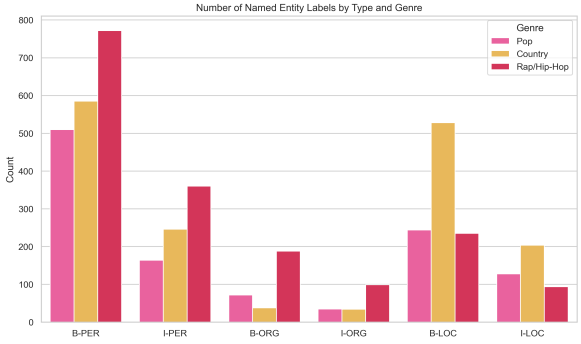

Figure 3: Number of named entity labels by type and genre.

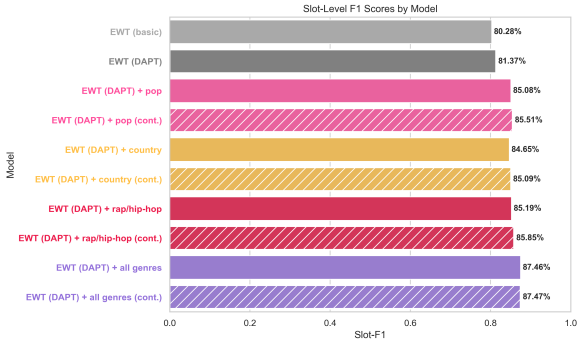

Figure 4: Slot-Level F1 score for each model.

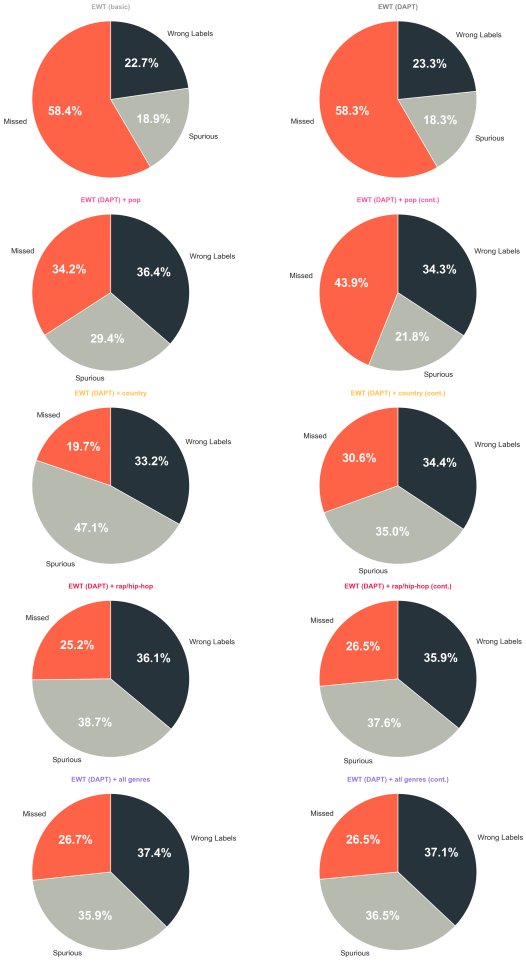

Figure 5: Percentage of each error type in all models.